RESEARCH BIODATA

NOOR LIDE BINTI ABU KASSIM

School of Educational Studies
Psychometrics & Education Evaluation – Standard Setting & Test Validation

Doctor of Philosophy (2007)

THESIS:

USING THE RASCH MEASUREMENT MODEL FOR STANDARD SETTING OF THE ENGLISH LANGUAGE PLACEMENT TEST AT THE IIUM

ABSTRACT:

With the use of cutscores and standards for making high-stakes educational decisions, efforts should be made to search for more defensible standard setting methods. This study represents an effort to this end. The main intent of this study is to investigate the efficacy of the Objective Standard Setting Method (OSS), which is based on the Rasch Measurement Model, in constructing multiple cutscores that are valid and defensible on tests utilizing diverse item types. The OSS, which was developed by Stone (1996) to set a single cutscore on tests utilizing selected-response items, has been demonstrated to yield valid results. However, its efficacy in handling other item types and the construction of multiple cutscores has yet to be empirically established. As the quality of the tests used in the standard setting endeavour influences the validity of derived cutscores as well as the validity of examinee classification, assessment-related issues are also of major concern. Measurement theory is one other aspect that requires serious consideration. The need for a measurement model that transforms counts correct into interval linear measures that are neither sample-bound nor test-bound, and at the same time references an examinee's performance (on the test) and status (based on the standards set) directly to the measured construct cannot be underrated. The same applies to the capacity to resolve important measurement and standard setting issues. In this study the efficacy of the OSS was examined in the context of the English Language Placement Test conducted at the IIUM. It was found that with the use of the OSS, multiple cutscores on diverse item types can be easily set without compromising the validity of the derived cutscores or standards. Additionally, with the use of the OSS, the desired level of attainment can be directly translated onto the measured construct and, thus, allowing the standards set to have real meaning and not just proportions of correct answers. The Rasch measurement model has also proved to be useful in resolving fundamental issues in measurement and standard setting. However, one cautionary word is necessary. Regardless of how sound a standard setting method is, the results of a standard setting study are bound to be impacted by test quality, judge

competency, performance level descriptions and other variables in the standard setting process. This has been demonstrated very clearly in this study. Steps must, therefore, be taken to address all these issues to ensure that the reliability and validity of derived cutscores and standards are not compromised.

PUBLICATIONS & CONFERENCES:

Conference Presentations:

Invited Paper:
Noor Lide Abu Kassim (2006). Using the Many-facet Rasch Model (FACETS) for the Construction of Cutscores and Resolving Standard Setting Issues, Pacific Rim Objective Measurement Symposium, 27 – 29 June, 2006, Hong Kong.

Paper Presentation:
Noor Lide Abu Kassim & Bond, T. G. (2006). Use of the Many-facet Rasch Model in Resolving Standard Setting Issues. International Objective Measurement Workshop (IOMW), 5-7 April 2006, University of California, Berkeley, USA.

USING THE RASCH MEASUREMENT MODEL FOR STANDARD SETTING OF THE ENGLISH LANGUAGE PLACEMENT TEST AT THE IIUM

NOOR LIDE BINTI ABU KASSIM

UNIVERSITI SAINS MALAYSIA

2007

5.7 RESULTS OF THE STANDARD SETTING STUDY

The following subsections present the (1) criterion points (i.e., the initial cutscores) established in the standard setting study, (2) final cutscores constructed after accounting for error of measurement (3) cutscores as applied to the relevant examinee distributions, (4) number and percentage of examinees categorized in each level of performance based on the final cutscores, and (5) the descriptive statistics of examinee distribution in each performance category.

5.7.1 Cutscores: Grammar Subtest

5.7.1.1 The Criterion Points and Final Cutscores

The criterion points for the grammar subtest established in the standard setting study are presented in Table 5.42. Along with the criterion points are the standard error, and the lower and upper boundaries of the criterion regions estimated with ± 1.6 standard errors of the calibrated item measures. Please note that the values for the standard errors have been rounded up to 2 decimal points for reporting purposes but not for the calculation of the ± 1.6 S.E. of the criterion points.

Table 5.42: Grammar Subtest Criterion Points Estimated With ± 1.6 Standard Errors

Criterion Point	Standard Error	-1.6 SEM	Measure	+1.6 SEM
4	0.05	+0.29	+0.37	+0.44
3	0.05	+0.05	+0.13	+0.20
2	0.05	-0.12	-0.04	+0.03
1	0.05	-0.17	-0.09	-0.02

Figure 5.35 gives a visual representation of the criterion region for each of the grammar criterion points. Note that the criterion regions for three of the four points overlap to varying extents (Criterion Points 1, 2 and 3) with one another. There is,

however, a small gap between the criterion regions for points 3 and 4. These overlaps indicate lack of a clear separation between the first three criterion points.

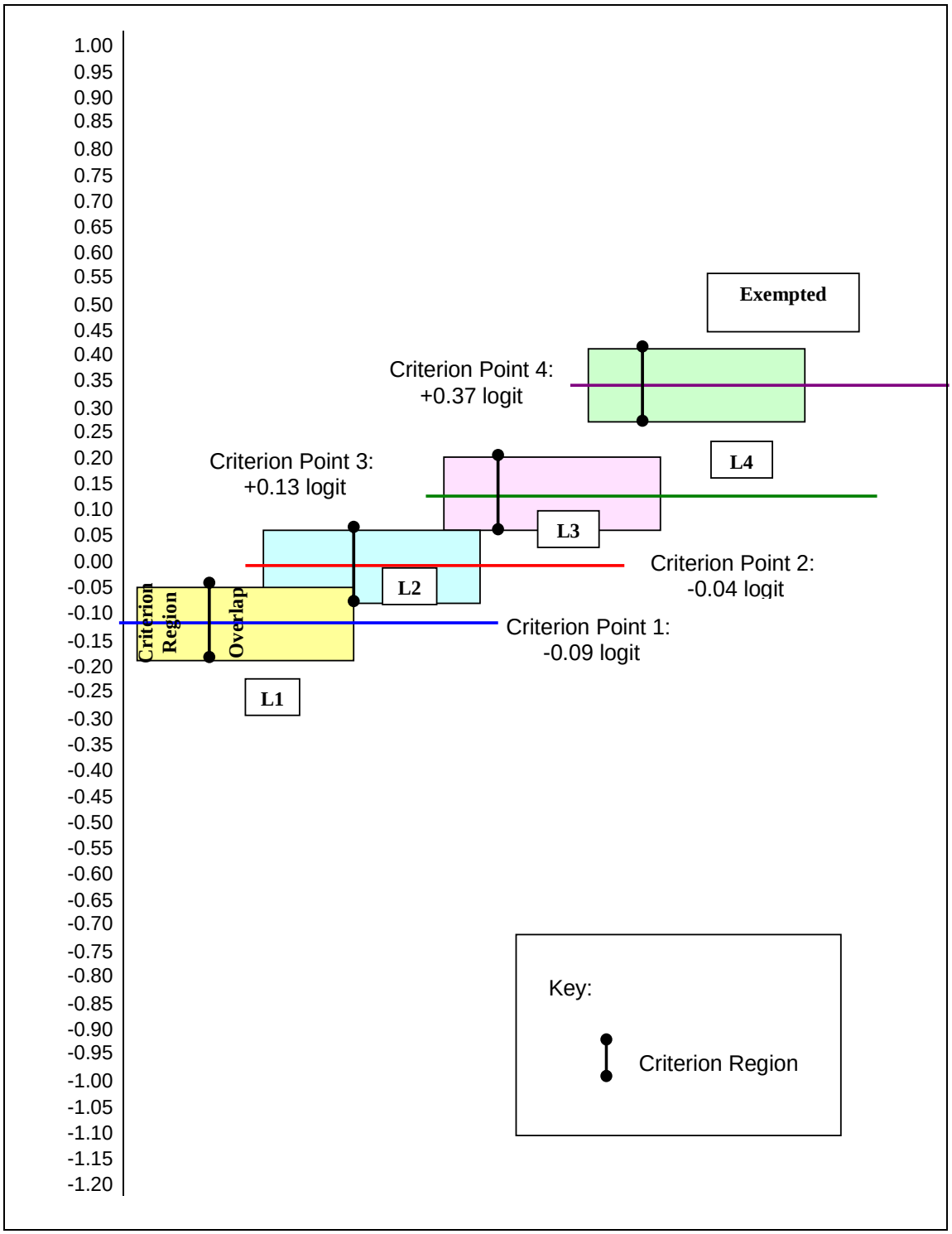

Figure 5.35: Grammar Subtest Criterion Points marked with ± 1.6 Standard Errors

In estimating the final cutscores, the criterion points determined by the standard setting judges were adjusted by adding +1.6 SEM to the criterion measures. As mentioned in Chapter 4, this decision was taken to ascertain that examinees have the required level of proficiency (i.e., guarantee quality) in order to be placed in a given proficiency level.

Table 5.43 presents the final cutscores with the addition of the confidence interval of approximately 95%. As expected the first and second cutscores are extremely close together with an almost negligible difference (0.05 logit). The distance between the second and the third cutscores (0.17 logit) is also very small. The difference between the third and the fourth cutscores, on the other hand, is slightly larger (0.24 logit) but on the whole, not large enough to comfortably separate the two levels of performance.

Table 5.43: Grammar Subtest Final Cutscores

Final Cutscore	Measure (logit)
Cutscore 4	+0.44
Cutscore 3	+0.20
Cutscore 2	+0.03
Cutscore 1	-0.02

5.7.1.2 Categorization of Examinees

Figure 5.36 graphically shows the close proximity of the cutscores as applied to the examinee distribution. It must be noted that this distribution of examinees based on the cutscores does not take into account the measurement error related to the calibration of examinee ability.

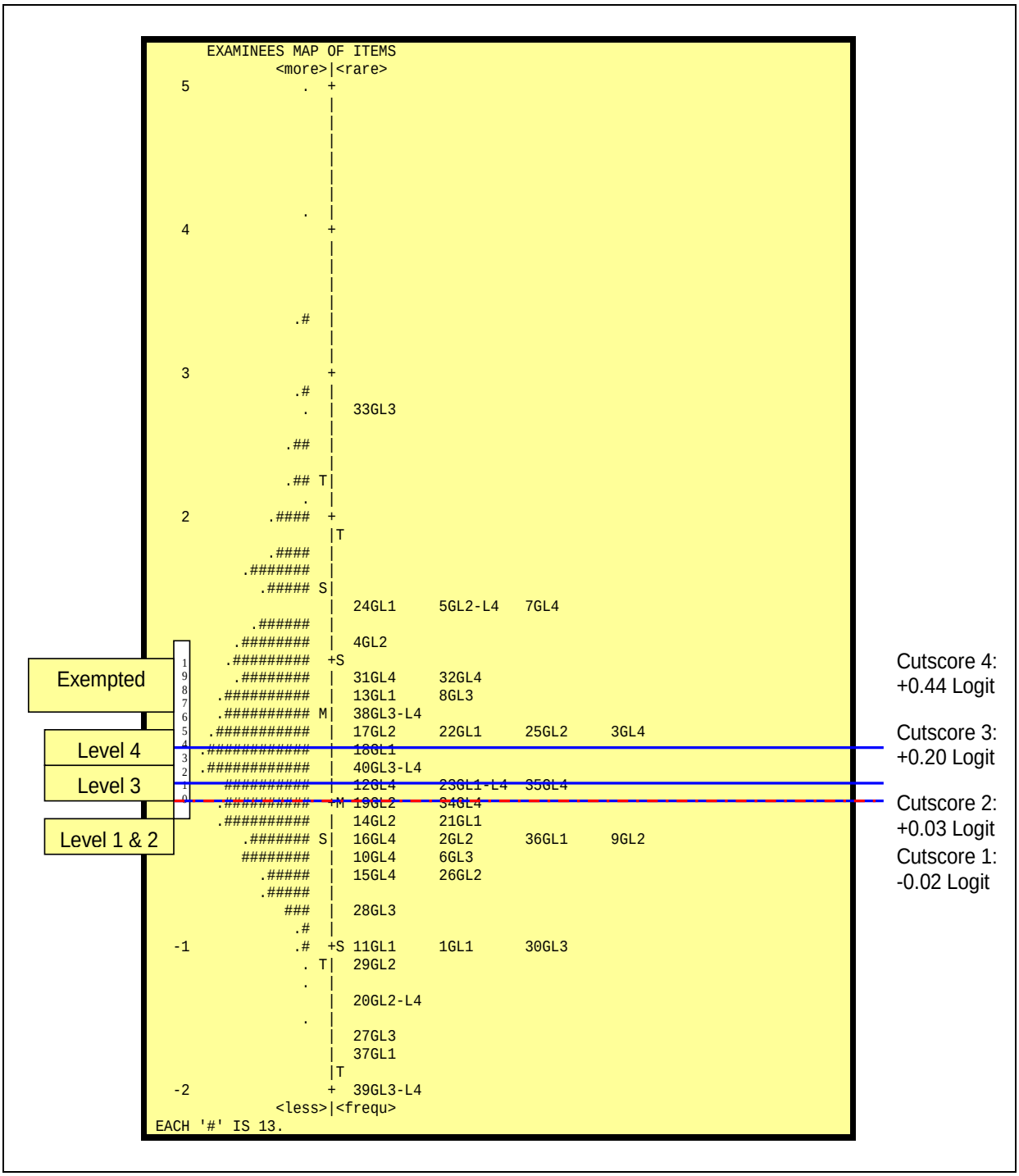

Figure 5.36: Grammar Subtest Final Cutscores Applied to Examinee and Item Distributions (Wright Map)

In Chapter 4, it is indicated that to ascertain a clear pass/fail decision at 95% confidence level, examinees' measures have to be adjusted by lowering or subtracting the examinee calibrated measures by 1.6 S.E. (Refer to Figure 5.37). This would give a 95% confidence level that examinee ability measures do not

overlap the criterion region and are located at or above the cutscore. These adjusted measures are the ones used in the classification of examinees into the respective performance levels or categories.

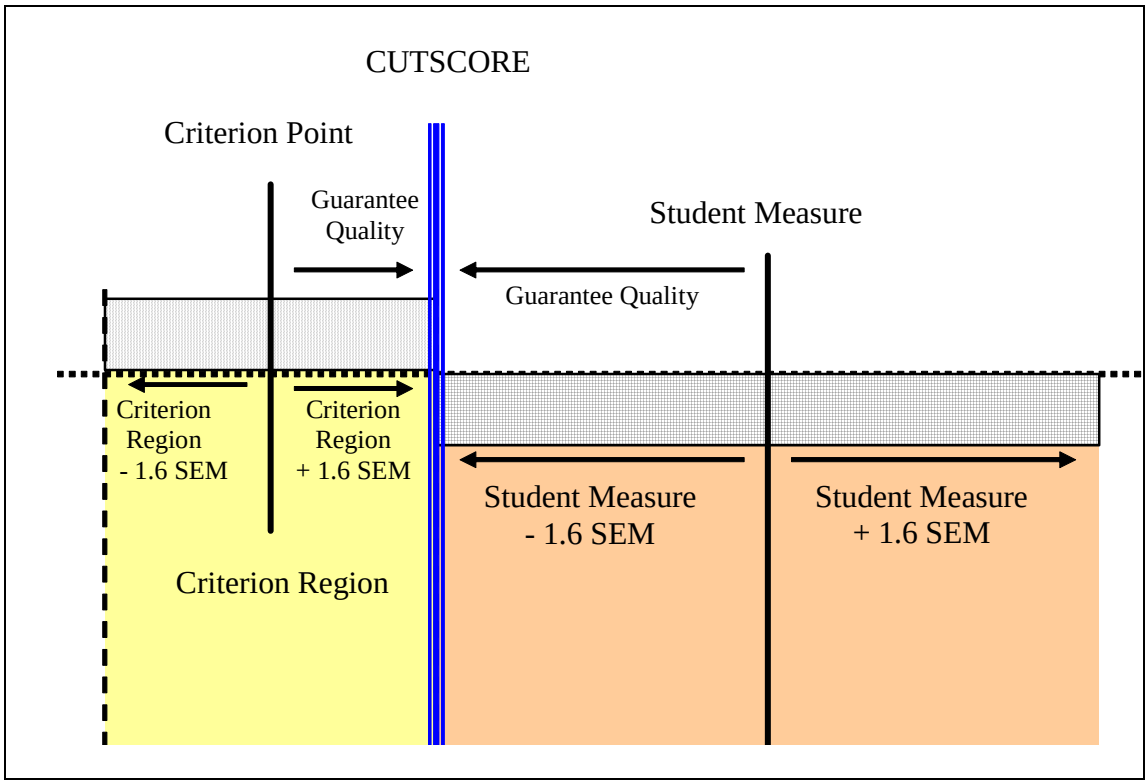

Figure 5.37: Adjusting Examinees' Calibrated Measures by -1.6 Standard Errors

Table 5.44 gives the frequency and percentage of examinees who are classified in the five categories. About 50% (*n* = 1309) of the examinees are classified in the lowest proficiency category. The 'Exempted' category has about 25% (*n* = 620) of the sample tested. Only ± 21% (*n* = 520) of the examinees are distributed between Levels 3 and 4 (11.4% and 9.9% respectively). Note that not a single examinee has been classified as belonging to Level 2. This is expected given the extremely small difference between the first and second cutscores.

Table 5.44: Frequency and Percentage of Examinees by Proficiency Level (Grammar Subtest)

Level	Frequency	Percentage
Exempted (+0.44 to Highest)	620	25.3
Level 4 (+0.20 to +0.43 logit)	242	9.9
Level 3 (+0.03 to +0.19 logit)	278	11.4
Level 2 (-0.02 to +0.02 logit)	0	0
Level 1 (Lowest to -0.01 logit)	1309	53.5

Table 5.45, on the other hand, presents the mean examinee ability estimates and the standard deviation of the estimates for each category. The mean ability estimates for the middle categories are not well-differentiated from one another. For example, the difference in mean ability for Levels 3 and 4 is 0.23 logit. The mean ability estimates that are reasonably differentiated based on the derived cutscores are the ones for Level 4 and the Exempted category (0.70 logits). This difference, however, is more the result of the large distribution of examinees in the Exempted category rather than the separation between the two cutscores.

Table 5.45: Mean Ability and Distribution Statistics by Proficiency Level (Grammar Subtest)

Level	N	Mean	Std. Deviation
Exempted	620	+1.03	0.47
Level 4	242	+0.33	0.07
Level 3	278	+0.10	0.06
Level 2	0	-	-
Level 1	1309	-0.60	0.40

5.7.2 Cutscores: Reading Subtest

5.7.2.1 The Criterion Points and Final Cutscores

The first and second criterion points for the reading subtest are quite differentiated with a difference of 0.6 logit (Table 5.46). Considering that examinee distribution spans about 3 logits, this difference is quite reasonable in measurement terms. The distance between the second and third criterion points (0.52 logit) also allows for a clear differentiation in examinee ability. The distance between the third and fourth criterion points, however, is very slight (0.08 logit). The lower and upper boundaries of the four criterion points estimated with ± 1.6 Standard Errors (i.e., criterion regions) are given in Table 5.46.

Table 5.46: Reading Subtest Criterion Points Estimated With ± 1.6 Standard Errors

Criterion Point	Standard Error	-1.6 SEM	Criterion Measure	+1.6 SEM
4	0.05	+0.31	+0.39	+0.47
3	0.05	+0.23	+0.31	+0.38
2	0.05	-0.29	-0.21	-0.14
1	0.05	-0.89	-0.81	-0.73

Figure 5.38 gives a visual representation of the criterion region for each of the reading criterion points. The figure indicates that the criterion regions for the first three points are well-differentiated. However, the close proximity of the top two criterion points, and the overlap between the criterion regions of the two criterion points show no clear separation between these two points.

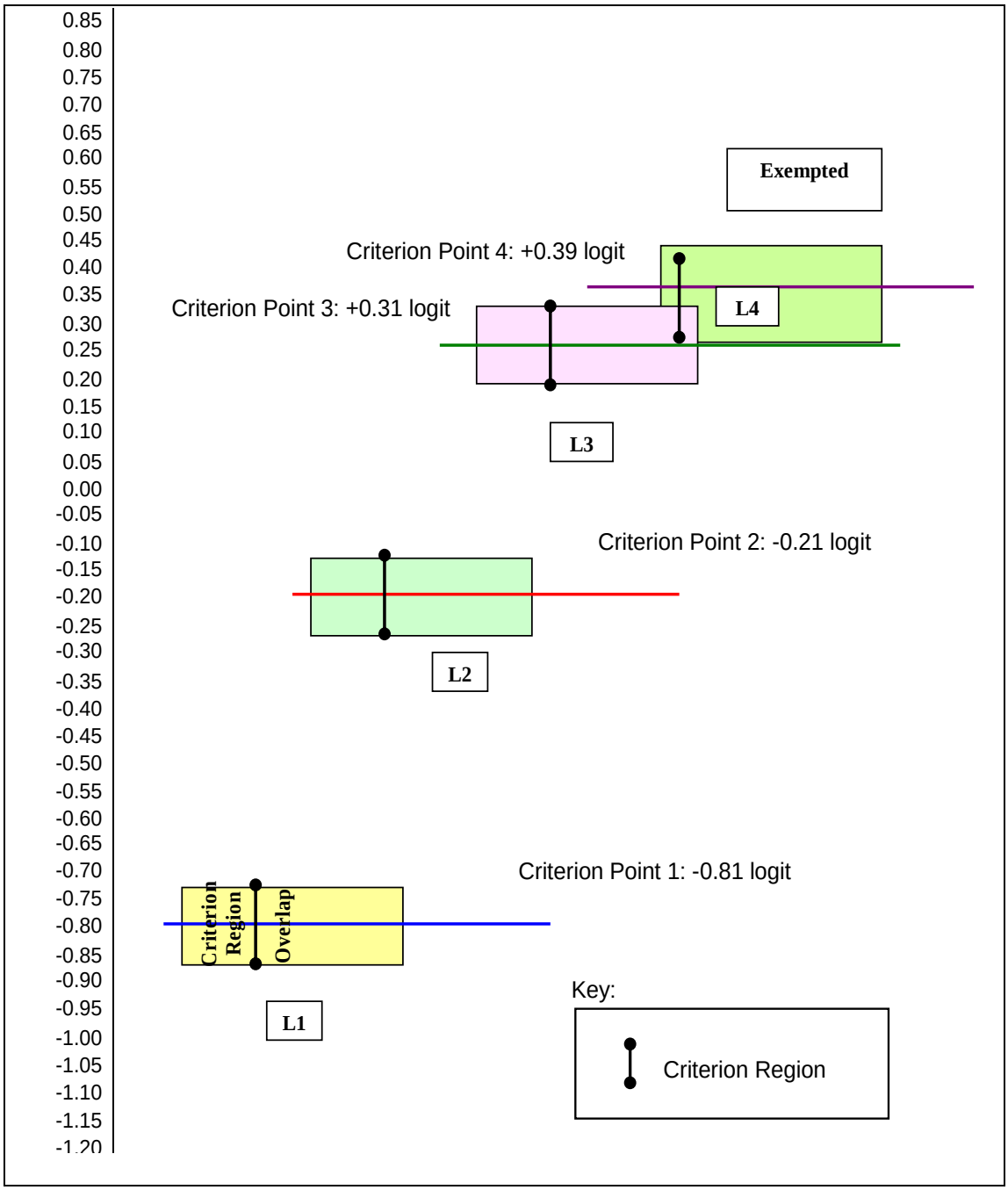

Figure 5.38: Reading Subtest Criterion Points Marked with ± 1.6 Standard Errors

Table 5.47 gives the final cutscores whereas the Wright map (Figure 5.39) shows the cutscores in relation to item locations and examinee distribution. From the map, quite a large number of examinees are expected to fall in Level 2, Level 3 and the 'Exempted' categories. A smaller number of examinees are expected to be classified in Level 1 and an even smaller number in Level 4.

Table 5.47: Reading Subtest Final Cutscores

Cutscore	Measure (logit)
Cutscore 4	+0.47
Cutscore 3	+0.38
Cutscore 2	-0.14
Cutscore 1	-0.73

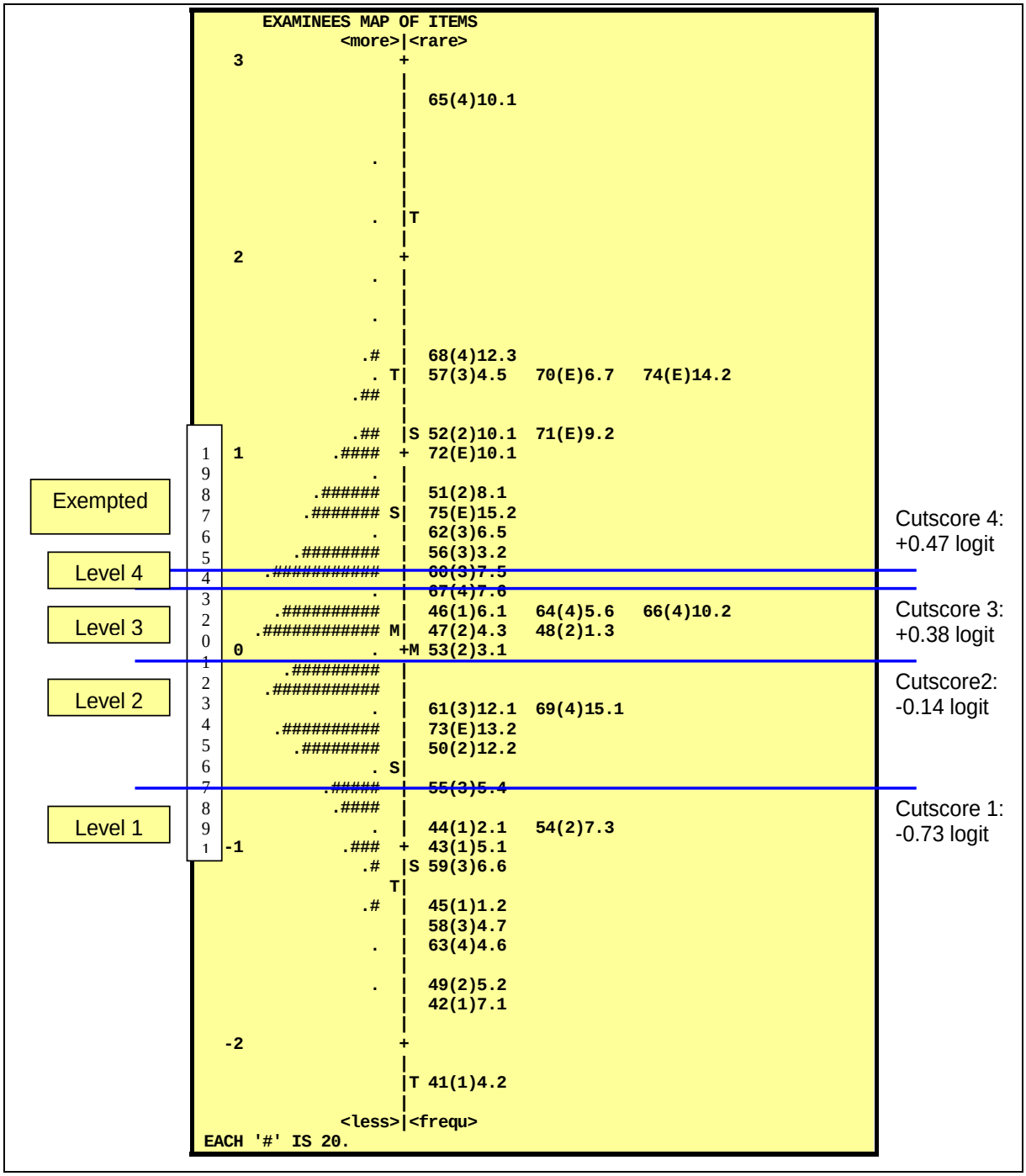

Figure 5.39: Reading Subtest Final Cutscores Applied to Examinee and Item Distributions (Wright Map)

5.7.2.2 Categorization of Examinees

Table 5.48 gives the frequency and percentage of examinees who fall in the five categories as classified by the cutscores. The largest number of examinees fall in the first level (*n* = 900). Level 2 has the second largest number of examinees (*n* = 856). This is followed by Level 3 (*n* = 515). The Exempted category has 178 examinees. Level 4, as a result of the close proximity of the third and fourth cutscores, has zero examinees.

Table 5.48: Frequency and Percentage of Examinees by Proficiency Level (Reading Subtest)

Level	Frequency	Percentage
Exempted (+0.47 to Highest)	178	7.3
Level 4 (+0.38 to +0.46 logit)	0	0.0
Level 3 (-0.14 to +0.37 logit)	515	21.0
Level 2 (-0.73 to -0.15 logit)	856	35.0
Level 1 (Lowest to -0.74 logit)	900	36.7

Table 5.49 gives the mean ability and the distribution of examinee ability estimates for each category. From the table, it is evident that the examinee mean ability estimates for the extreme categories are more differentiated than the examinee mean ability estimates for the middle categories.

Table 5.49: Mean ability and Distribution Statistics by Proficiency Level (Reading Subtest)

Level	N	Mean	Std. Deviation
Exempted	178	0.72	0.27
Level 4	0	-	-
Level 3	515	0.07	0.15
Level 2	856	-0.46	0.16
Level 1	900	-1.17	0.33

5.7.3 Cut Scores: Writing Subtest

5.7.3.1 Criterion Points and Final Cutscores

In calculating the cutscores for the writing subtest, ratings given by the standard setting judges were recoded to the corresponding Rasch-Thurstone thresholds (.5 cumulative probabilities) derived from the Facets analysis of examinees' essay ratings on the five items or essay criteria. These threshold estimates were then averaged to get the mean estimates for each judge. To calculate the criterion measure for each criterion point, individual judges' estimates for the individual criterion point were again averaged.

Table 5.50 presents the measures for the respective criterion points, the standard errors of the calibrated items, and the estimates for the upper and lower boundaries of the criterion regions. Figure 5.40 gives a graphic representation of the information delineated in Table 5.50. Note that the error estimate for the fourth criterion point is considerably larger than for the ones for the second and third criterion points. This is the result of the small observations in the extreme categories of the rating scale.

Table 5.50: Writing Subtest Criterion Points with ± 1.6 Standard Errors

Criterion Point	Standard Error	-1.6 SEM	Criterion Measure	+1.6 SEM
4	0.31	3.68	+4.17	4.66
3	0.08	1.18	+1.31	1.44
2	0.05	-1.11	-1.03	-0.95
1	0.13	-4.48	-4.28	-4.08

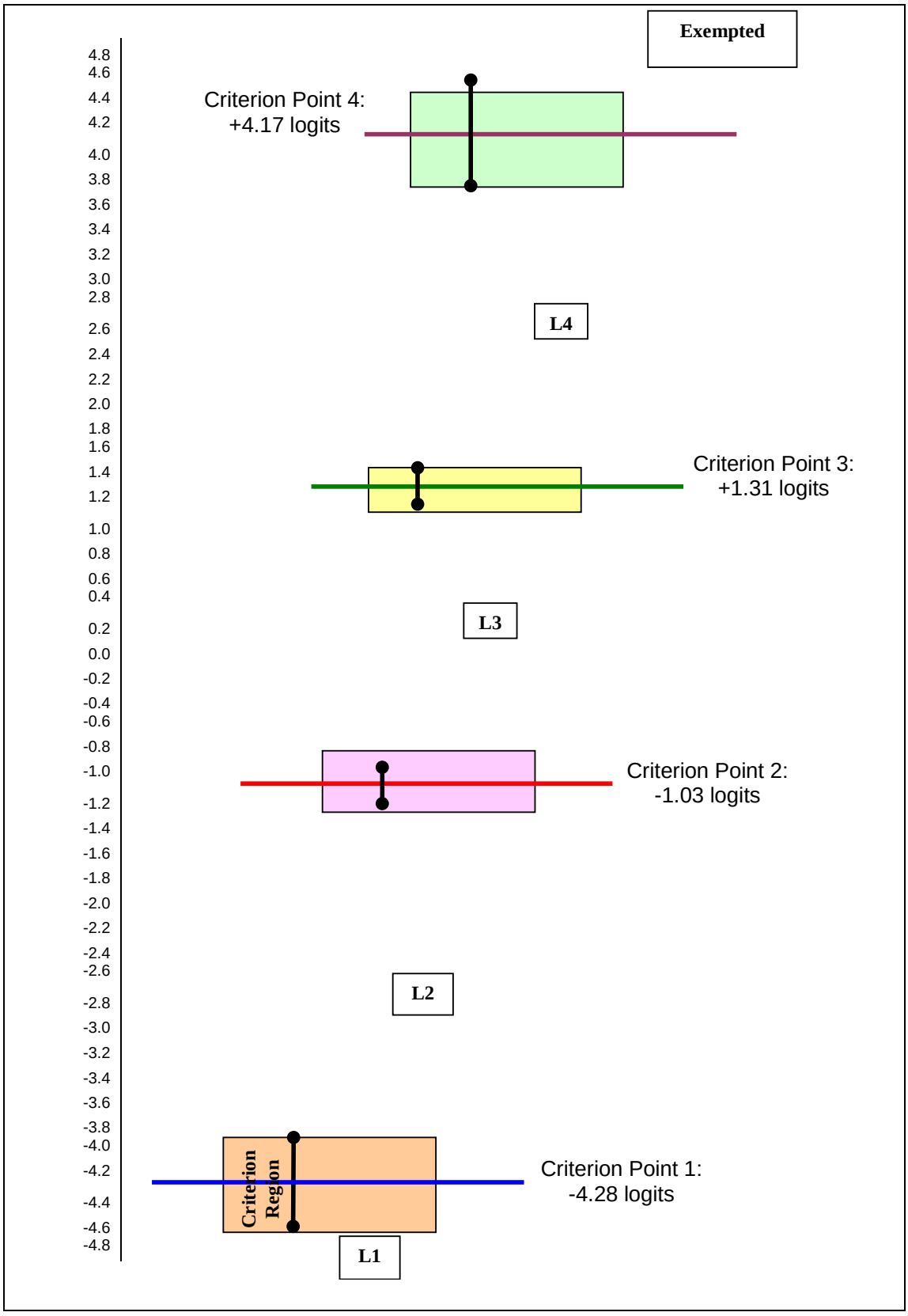

Figure 5.40: Writing Subtest Criterion Points Marked With ± 1.6 Standard Errors

Table 5.51 gives the final cutscores for the writing subtest. Figure 5.41, on the other hand, presents the cutscores as applied onto the Wright map of the writing subtest. The Wright map indicates that the cutscores are well-separated and spread across the examinee distribution. However, the two cutscores at the extreme ends have been either over-estimated or under-estimated by the standard setting judges. The first cutscore is located too close to the lower end of the distribution whereas the fourth cutscore is located close to the uppermost end. This would inevitably affect the number of examinees classified in the top and bottom categories.

The distance between the first and second cutscores is quite substantial and more than half of the measured examinees are located between these two cutscores. The distance between the second and third cutscores, on the other hand, is less than the distance between the first two. The number of examinees located between these two cutscores is smaller but still substantial. From the Wright map it can be seen that the number of examinees classified in Level 1 is highly likely to be very small. The number of examinees above the Level 4 cutscore is expected to be even smaller.

Table 5.51: Writing Subtest Final Cutscores

Cutscore	Measure (logit)
Cutscore 4	4.66
Cutscore 3	1.44
Cutscore 2	-0.95
Cutscore 1	-4.08

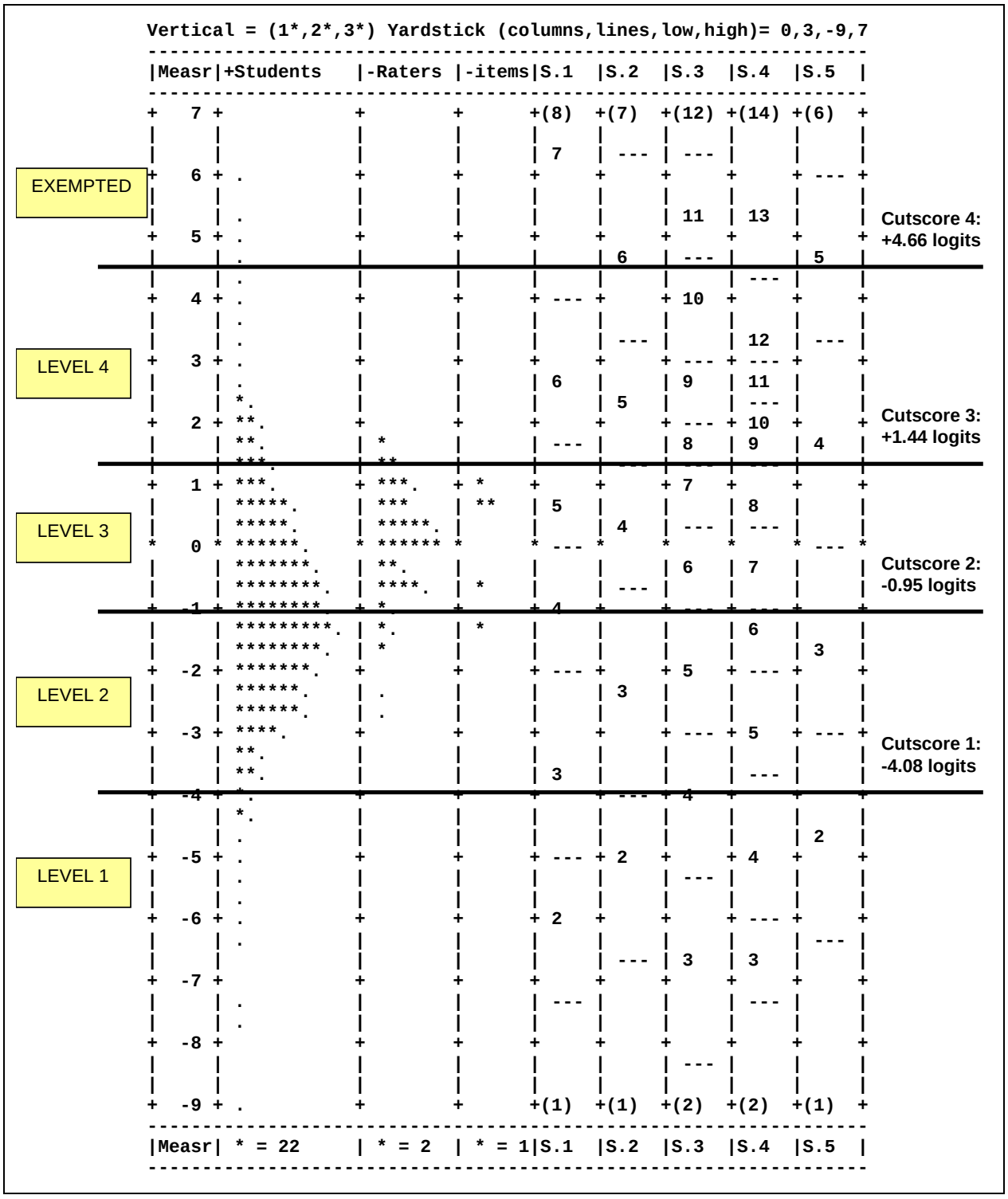

Figure 5.41: Writing Subtest Final Cutscores Applied to Examinee and Item Distributions

5.7.3.2 Categorization of Examinees

Table 5.52 gives the frequency and percentage of examinees who fall in the five categories. Only 1 examinee is classified in the 'Exempted' category. About 12% (*n* = 289) of the examinees are categorized in the lowest proficiency category. Level 2 has the largest percentage of examinees (60.7%) as the first and second cutscores cut across more than half of the total number of examinees measured.

Table 5.52: Frequency and Percentage of Examinees by Proficiency Level (Writing Subtest)

Level	Frequency	Percentage
Exempted (+4.66 to Highest)	1	0.0
Level 4 (+1.44 to +4.65 logit)	77	3.1
Level 3 (-0.95 to +1.43 logit)	595	24.3
Level 2 (-4.08 to -0.96 logit)	1486	60.7
Level 1 (Lowest to -4.09 logit)	289	11.8

Table 5.53 presents the examinee mean ability estimates and the distribution statistics of these estimates for each of the performance categories. The examinee mean ability estimates for all the categories are well-differentiated and they range from about 2 to 3 logits. Please note that there is one missing examinee as his/her ability was unmeasurable due to extreme scores (zeros on all items).

Table 5.53: Mean Ability and Distribution Statistics by Proficiency Level (Writing Subtest)

Level	N	Mean	Std. Deviation
Exempted	1	4.96	-
Level 4	77	2.31	0.70
Level 3	591	-0.01	0.66
Level 2	1485	-2.40	0.83
Level 1	289	-4.96	0.88

5.7.4 Cutscores: Compensatory Approach

5.7.4.1 Criterion Points and Final Cutscores

The results of the compensatory approach indicate somewhat reasonable criterion points given the examinee distribution. The difference in measures for the criterion points range from about 0.5 logit to 0.9 logit. The smallest difference is between the second and third criterion points whereas the largest difference is between the third and fourth criterion points (Table 5.54).

The standard error of measurement for the first criterion point is the largest (0.14 logit). This again is influenced by the small number of observations in the low categories of the essay criteria. Standard errors for the middle categories are much smaller as more observations are distributed in the middle rating categories. Figure 5.42 gives a graphic representation of the criterion points and the criterion regions.

Table 5.54: Criterion Points Estimated With ± 1.6 Standard Errors (Compensatory Approach)

Criterion Point	Standard Error	-1.6 SEM	Criterion Measure	+1.6 SEM
4	0.08	+1.34	+1.46	+1.58
3	0.05	+0.45	+0.53	+0.61
2	0.06	-0.05	+0.04	+0.13
1	0.14	-0.74	-0.52	-0.30

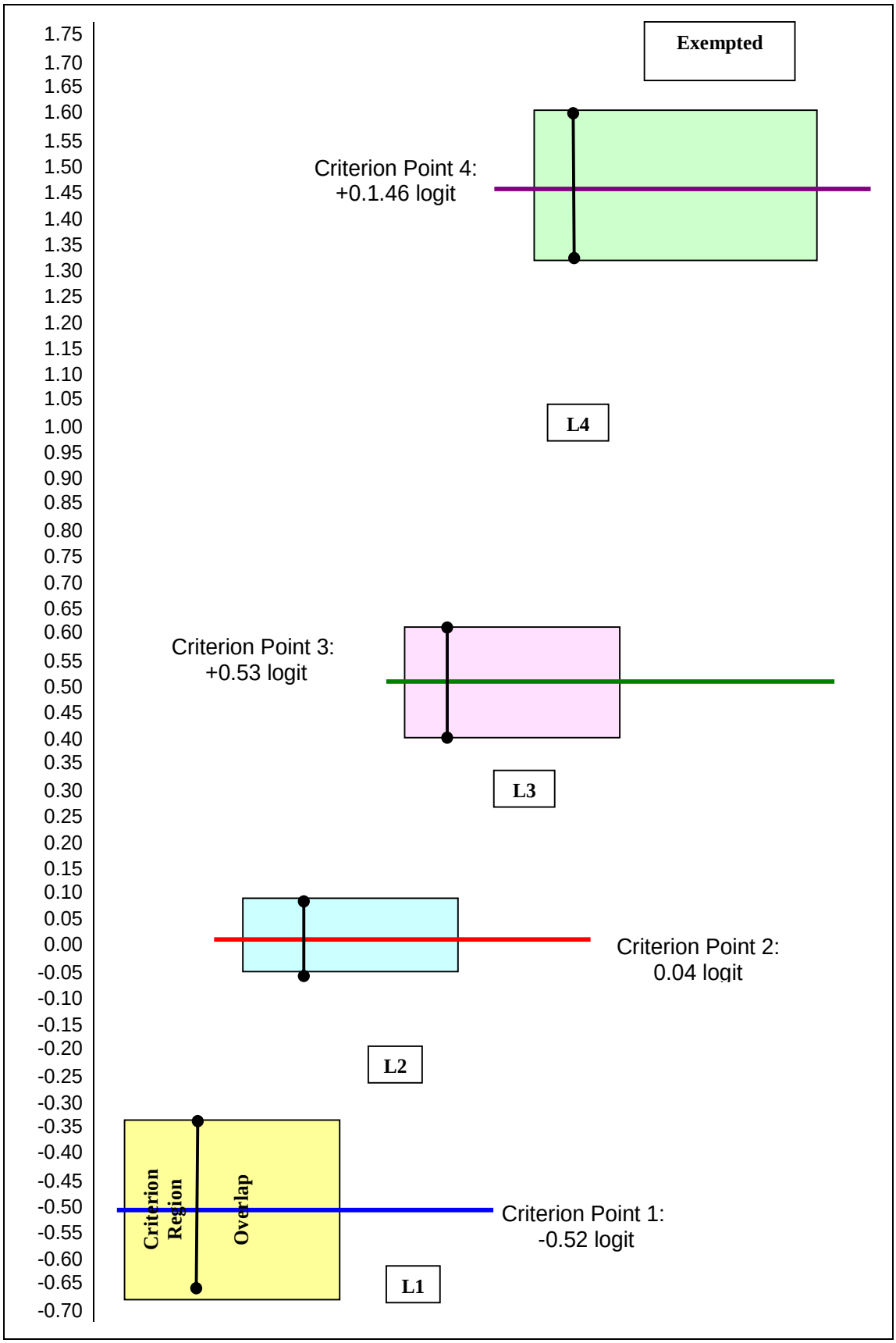

Figure 5.42: Criterion Points Marked with ± 1.6 Standard Errors
(Compensatory Approach)

The final cutscores for the compensatory approach are presented in Table 5.55. The cutscores as applied to the examinee distribution are given in Figure 5.43. It is evident that there is clear separation between the four cutscores to yield appropriate classifications of examinees. Nevertheless, given the small number of examinees at the top end of the examinee distribution, and the somewhat over-estimation of the fourth cutscore, a small number of examinees are expected to fall in the Exempted category.

Table 5.55: Final Cutscores (Compensatory Approach)

Cutscore	Measure (logit)
Cutscore 4	+1.58
Cutscore 3	+0.61
Cutscore 2	+0.13
Cutscore 1	-0.30

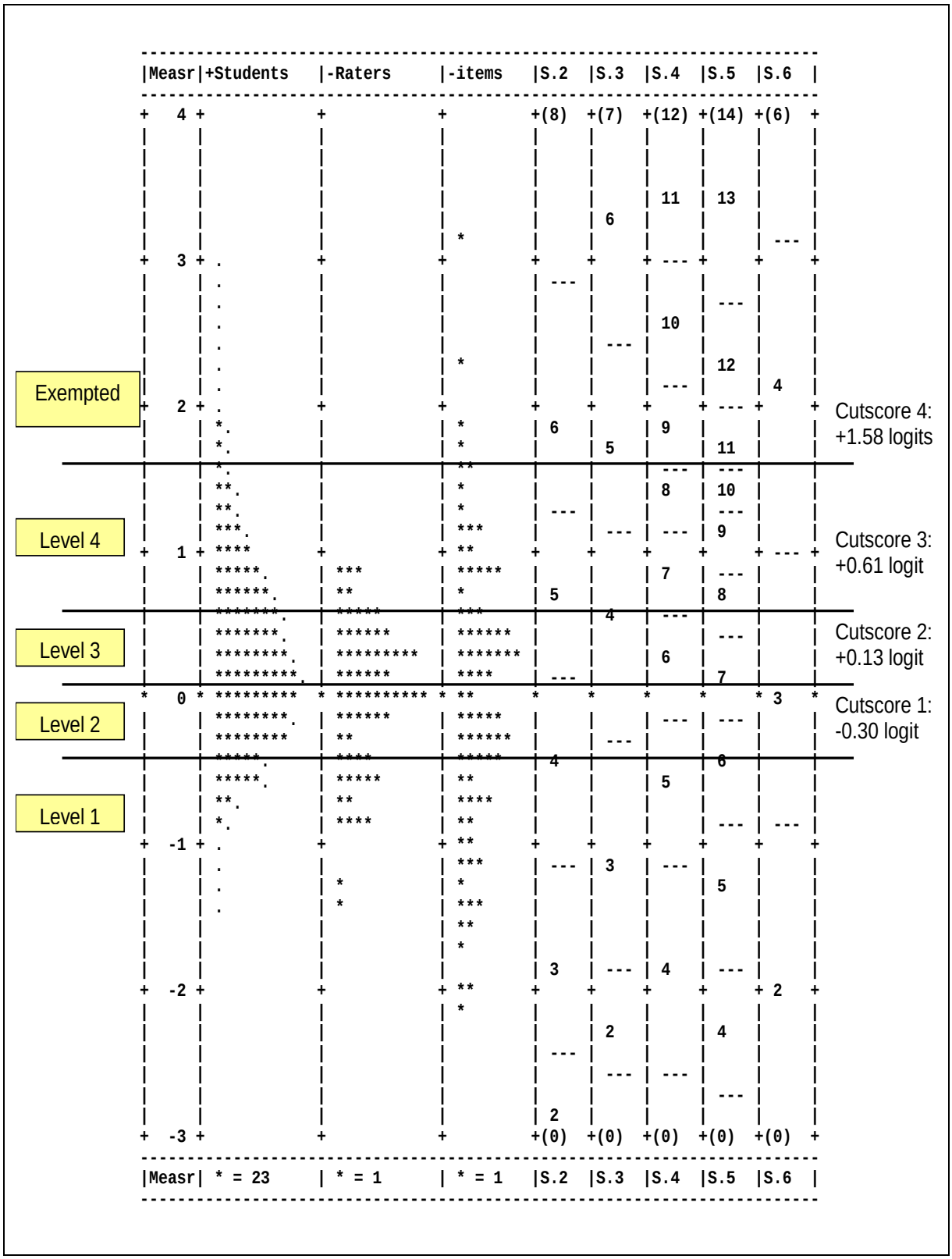

Figure 5.43: Final Cutscores Applied to Examinee and Item Distributions (Compensatory Approach)

5.7.4.2 Categorization of Examinees

The distribution of examinees by each performance category reveals that only 42 examinees (1.7%) are classified in the Exempted category. The bulk of examinees are distributed in the lowest category (*n* = 971). The other categories have rather reasonable distributions of examinees (Table 5.56).

Table 5.56: Frequency and Percentage of Examinees by Proficiency Level (Compensatory Approach)

Level	Frequency	Percentage
Exempted (+1.58 to Highest)	42	1.7
Level 4 (+0.61 to +0.1.57 logit)	348	14.2
Level 3 (+0.13 to +0.60 logit)	498	20.3
Level 2 (-0.30 to +0.12 logit)	590	24.1
Level 1 (Lowest to -0.29 logit)	971	39.6

Table 5.57, on the other hand, shows the examinee mean ability estimates and the distribution statistics for each performance category. The top two categories are quite well-differentiated in terms of ability and they also indicate a relatively large spread in examinee distribution.

Table 5.57: Mean Ability and Distribution Statistics by Proficiency Level (Compensatory Approach)

Level	N	Mean	Std. Deviation
Exempted	41	1.94	0.25
Level 4	342	0.98	0.26
Level 3	494	0.35	0.13
Level 2	584	-0.10	0.12
Level 1	971	-0.69	0.28

5.8 EFFICACY OF THE OSS

In this study, the OSS was applied to SR and CR items, and a mix of these two item types to produce multiple cutscores for the purposes of the EPT. Therefore, the evaluation of the efficacy of the OSS focuses on evidence to demonstrate the procedural, internal and external validity of the OSS-derived cutscores in relation to these issues. Sources of evidence for these aspects of validity are presented in the following subsections.

5.8.1 Procedural Validity

The evidence accrued for procedural validity centres on two sources. The first source of evidence pertains to the quality of the implementation of the standard setting study and the second relates to the appropriateness of the procedure used in the setting of the cutscores (Kane, 1994). The first aspect is investigated by examining the adequacy of judge selection and training, and the adequacy of data collection procedures.

Appropriateness of the standard setting procedure, on the other hand, is examined by looking at judges' ability to perform the judgment task. This centres on questions related to judge expertise, judges' ability to identify essential items, judges' confidence in the selection of essential items, judges' confidence in the classification of examinees and judges' views on the judgment task (i.e., the standard setting procedure).

From judges' feedback on the standard setting study, other issues related to standard setting and assessment have also surfaced. These include matters pertaining to item writing and the rating scale used in the assessment of examinees' essays. As these issues are of immediate relevance to assessment and standard setting they are also reported here.

5.8.1.1 Implementation of the Standard Setting Study

5.8.1.1.1 Judge Selection

For cutscores and standards to be credible, a standard setting study should involve not only a relatively large number of judges with relevant background but also judges who are competent in making the decisions that they are expected to make. In this study, the selection of judges could not be carried out as intended by the researcher. However, the pool of judges used had the necessary background in the teaching and learning of English as a second language at the Centre as they had been with CELPAD for at least two years and had been teaching the language support courses for the duration of their service. It is also important to note that these judges were item writers for the English language placement test; some of them were fairly experienced whereas others were newly-recruited. These judges, therefore, had varying amounts of exposure to item writing and the item specifications of the EPT.

5.8.1.1.2 Judge Training

Table 5.58 gives judges' feedback on the success of the training given prior to the standard setting study. Of the 20 judges who responded to this question, 11 (55%) indicated that the training was successful, whereas 8 (40%) of the judges felt that it was only partially successful. Only one judge (5%) indicated that the training was not successful.

Table 5.58: Success of Judge Training

	Frequency	Percentage (%)
Not successful	1	5.0
Partially Successful	8	40.0
Successful	11	55.0
Very Successful	0	-

Some comments have also been given with regard to training. Several judges (*n* = 3) wanted more training and more examples of test items; another 3 wanted more hands-on practice while one judge wanted a longer training session.

5.8.1.1.3 Procedures for Data Collection: Time Allocation

In terms of time given for the judging of essential items about half of the judges across all the subtests agreed that the amount of time allocated was sufficient (Table 5.59). However, 50% of the judges felt that the time given for the reading subtest was too little. This is somewhat similar to what they felt about the grammar subtest. Given the number of items on the grammar subtest (40 items) and the reading subtest (35 items and 5 reading passages) this is quite understandable. On the other hand, about 25% of the judges felt that the time allocated for the writing subtest was insufficient. For this subtest, judges only had to rate the required performance on five essay items. There are 20% missing responses for the writing subtest.

Table 5.59: Adequacy of Time Allocation

Subtest	About Right	Too Little	Too Much	Missing
Grammar	55% (11)	40% (8)	-	5% (1)
Reading	50% (10)	50% (10)	-	-
Writing	55% (11)	25% (5)	-	20% (4)

5.8.1.1.4 Procedures for Data Collection: Adequacy of Performance Level Descriptions

In this standard setting study, performance level descriptions that describe the four cutscores were given to guide judges in the selection of essential items. Therefore, an important issue is the adequacy of the performance level descriptions in assisting the judges to perform the judgment task.

Overall, 56% of the judges considered that the performance level descriptions were successful in describing the desired cutscores (Table 5.60). On the other hand, 35% of the judges considered the descriptions only partially successful. However, none of the standard setting judges thought that the performance level descriptions were unsuccessful.

Table 5.60: Adequacy of Performance Level Descriptions

	Frequency	Percentage (%)
Not Successful	0	0
Partially Successful	7	35
Successful	13	56
Very Successful	0	-

Judges' opinion of the performance level descriptions was also elicited for each of the subtests. As far as the grammar subtest is concerned, a larger percentage (about 60% to 70%) of the judges felt that the performance level descriptions are adequate. This is particularly so for the top two cutscores.

Table 5.61: Adequacy of Performance Level Descriptions (Grammar)

Cutscore	Totally Adequate	Adequate	Partially Adequate	Totally Inadequate
Cutscore 1	-	60.0% (12)	40.4% (8)	-
Cutscore 2	-	65.0% (13)	35.0% (7)	-
Cutscore 3	-	70.0% (14)	30.0% (6)	-
Cutscore 4	-	70.0% (14)	30.0% (6)	-

Several comments on the adequacy of the performance level descriptions for the grammar subtest were also given; these are presented below. Clearly, there are judges who felt that the performance level descriptions were inadequate and, therefore, should be revised.

Judge 3:

"List of forms and structures occasionally incomplete"

"Incorrect terminology"

Judge 7:

"Some items may be difficult for Level 1 students (could it be possible that expectations are too high)".

Judge 14:

"Specs too long …Can they be condensed?"

The reading subtest elicited a somewhat similar response with regard to the first three cutscores (Table 5.62). More judges found the performance level descriptions for the fourth cutscore to be more adequate (73.7%) than the first three cutscores. One judge, on the other hand, felt the performance level description for Cutscore 1 to be inadequate.

Table 5.62: Adequacy of Performance Level Descriptions (Reading)

Cutscore	Totally Adequate	Adequate	Partially Adequate	Totally Inadequate
Cutscore 1	-	63.2% (12)	31.6% (6)	5.3% (1)
Cutscore 2	-	65.0% (13)	35.0% (7)	-
Cutscore 3	-	65.0% (13)	35.0% (7)	-
Cutscore 4	5.3% (1)	73.7% (14)	21.1% (4)	-

Judges' opinion of the first and second performance level descriptions for the writing subtest is not quite favourable (Table 5.63). Only 50 to 56 % of the judges found the descriptions to be adequate whereas about 44 % found the descriptions to be partially adequate for the two lower cutscores. Judges found the two upper

cutscores to be more adequate. The most adequate description is for the fourth cutscore (77.8%).

Table 5.63: Adequacy of Performance Level Descriptions (Writing)

Cutscore	Totally Adequate	Adequate	Partially Adequate	Totally Inadequate
Cutscore 1	-	50.0% (9)	44.4% (8)	5.6% (1)
Cutscore 2	-	55.6% (10)	44.4% (8)	-
Cutscore 3	-	66.7% (12)	33.3% (6)	-
Cutscore 4	-	77.8% (14)	22.2% (4)	-

Below are some of the comments given by the judges on the adequacy of the performance level descriptions of the writing subtest. It is evident that there are some disparities. Judges 5 and 8 did not face any problems with the descriptions; Judges 3 and 15 felt otherwise.

Judge 5:

"Not really. I thought the performance levels was more clear cut."

Judge 8:

"Not really. Just follow the descriptors."

Judge 3:

"Re-consider scores and corresponding descriptors?"

Judge 15:

"Descriptors seem to overlap"

5.8.1.2 Appropriateness of Standard Setting Procedure

A major concern in the evaluation of a standard setting procedure relates to the judgment task. In this study, appropriateness of the standard setting procedure is evaluated by eliciting information on judge expertise, judges' confidence in the selection of items, judges' confidence in the resulting standards, and judges' confidence in the standard setting method used.

5.8.1.2.1 Judge expertise

From the judges' responses to the open-ended questions in the evaluation form, it was found that some of the judges lacked a clear understanding of what the test items were testing (Judges 2, 5, 6, 7, 8, and 20). Examples of the comments that were given are as follows:

Judge 5:

"Maybe there should be a guideline/workshop on how to identify/choose items to be tested"

"I wasn't sure if the item I chose is the right one for the level of performance to be tested (Whether it really tests the particular level it's supposed to)"

"*I think more explanations or examples should be given in the test specifications so that everybody would have the same understanding/ idea of what the item really is – to avoid misinterpretation of the test item)"*

Judge 6:

"Determining the level of difficulty, esp. between Level 2 and 3. of items to be tested. It seems there's some overlapping between these two levels. It ends up with the items too difficult or too easy for the students"

Judge 7:

"Understanding the language of the test specifications (esp. reading) is especially challenging → we need more time, more knowledge, more hands-on practice working on understanding the test spec. There should be an example of the 'form' of the question. E.g. what does a question that test communicative value of a sentence look like?"

Judge 8:

"It takes time to understand the test specs"

Judge 20:

"Unsure of what the blueprint items really meant if seen as individual components and at different levels"

Judges also indicated that they require more exposure and explanation to help them perform the judgment task (Judges 2, 5, 7, 9, and 20). Examples of the comments given are stated below:

Judge 2:

"A longer duration of time should be allocated to the practical, hands-on activities so that a wider exposure to question structures is given."

Judge 5:

"...more explanations or examples should be given in the test specifications"

Judge 9:

"Thorough experience; more reading; lack of experience"

5.8.1.2.2 Identification of Essential Items

None of the judges felt that the identification of essential items was unsuccessful. However, only 50% of the judges indicated that it was successful (Table 5.64). The other 50% thought that they were only partially successful in identifying the essential items. Two of the judges did not respond to this question.

Table 5.64: Identification of Essential Items

	Frequency	Percentage (%)
Not Successful	0	0
Partially Successful	10	50
Successful	10	50
Very Successful	0	-

5.8.1.2.3 Confidence in the Selection of Essential Items

As regards the level of confidence in deciding the essentiality of items, only 10.5% of the judges were very confident (Table 5.65). A substantial percentage of the judges (63.2%) felt somewhat confident. The rest of the judges (26.3%) were confident of their decisions.

Table 5.65: Confidence in Deciding the Essentiality of Items

Level of Confidence	Frequency	Percentage (%)
Not confident	-	-
Somewhat confident	12	63.2
Confident	5	26.3
Very confident	2	10.5

With respect to the individual subtests, the highest level of confidence is for the grammar subtest (Very High, 5.0%; High, 55.0%) (see Table 5.66). The least amount of confidence is for the reading subtest (Low, 5%; Medium, 60%).

Table 5.66: Confidence in Deciding Essentiality of Items by Subtest

Subtest	Low	Medium	High	Very High
Grammar	-	40.0% (8)	55.0% (11)	5.0% (1)
Reading	5.0% (1)	60.0% (12)	30.0% (6)	5.0% (1)
Writing	-	47.4% (9)	42.1% (8)	10.5% (2)

When asked what problems they faced in judging the essentiality of items some judges gave the following comments:

Judge 5:

"I wasn't sure if the item I chose is the right one for the level of performance to be tested (Whether it really tests the particular level it's supposed to)"

"Sometimes I wasn't sure of the test specifications myself. Didn't understand how to test a particular item-maybe because didn't understand the item myself."

Judge 8:

"The same items but for different levels"

Judge 9:

"Too many items; some items are difficult to place according to level."

Judge 10:

"A little problem. To differentiate the reading passages according to different level (1-4)."

Judge 11:

"Especially in identifying the suitable items for each level"

Judge 20:

"Unsure of what the blueprint items really meant if seen as individual components and at different levels. Insufficient examples for guidance. Items are unclear and rather confusing."

From the comments given, the difficulty that judges faced in carrying out the judgment task has more to do with their understanding of the test specifications, what the test items are testing and identifying the suitable level of performance rather than the conduct of the judgment task.

5.8.1.2.4 Confidence in the Classification of Examinees

A somewhat different result is found in relation to judges' confidence in the classification of examinees based on the four cutscores (Table 5.67). Overall, the judges indicated high levels of confidence in examinee classification. Nonetheless, they expressed less confidence in the classification of examinees based on the middle cutscores than classification based on the top and bottom cutscores.

Table 5.67: Confidence in Classification of Examinees

Cutscore	Low	Medium	High	Very High
Cutscore 1	-	25.0% (5)	65.0% (13)	10.0% (2)
Cutscore 2	-	45.0% (9)	45.0% (9)	10.2% (2)
Cutscore 3	-	50.0% (10)	40.0% (8)	10.2% (2)
Cutscore 4	-	20.0% (4)	70.0% (14)	10.2% (2)

5.8.1.2.5 Confidence in the Standard Setting Method

As regards confidence in the standard setting procedure, only one judge (5%) felt very confident whereas 8 (40%) of the judges felt confident. The other 11 judges (55%) only felt somewhat confident of the standard setting method.

Table 5.68: Efficacy of the Standard Setting Method

Level of Confidence	Frequency	Percentage (%)
Not confident at all	0	-
Somewhat confident	11	55
Confident	8	40
Very confident	1	5

5.8.1.3 Other Issues

Three important issues have surfaced from the judges' feedback. The first relates to item writing; the second to the rating scale used; and the third concerns the standards expected and agreement between judges. The comments related to item writing are reported below:

Judge 6:

"Forming questions appropriate to the level".
"Distinguishing the inferencing/main idea/etc"

Judge 11:

"Constructing questions that fit the test specs"

Judge 12:

"Difficulty in writing options for testing tenses"
"Appropriacy of items to target level"
"Items may not come out the way we want it"

Judge 14:

"Making the grammar items to suit the EXACT requirement of the Blue Print"

In relation to the rating scale used, the concern was on differences in interpretation of the scale and the range of marks allocated for each of the essay items.

Judge 6:

"The language part because the range is too wide (esp. between average, good and fair)".
"The vocabulary section (esp. between level 2 and 3). The range is also quite wide".

Judge 7:

"Range of marks for average to good is very wide".

"Writing scores are highly subjective→sometimes it's difficult to decide exactly what is 'average', how many mistakes/errors would be considered 'few/many' what does 'does not obscure meaning' mean? I think the writing profile needs to be improved → what are the alternatives?"

Judge 12:

"Different interpretation of marking scheme".

The following comments reflect the judges' uncertainty about the expected standard and their concern for agreement with other judges.

Judge 7:

"There is a tendency to allocate marks on the lower end to avoid being labeled 'an easy grader' yet at the same time one wants to give students due credit."

Judge 12:

"Tend to remind ourselves about what others would think about the marks we give. (e.g. not too high etc)"
"What is standard and what is not standard?"

5.9 Internal Validity

Several sources of evidence of internal validity were investigated. The first relates to variability in judges' ratings of essential items to represent each criterion point (i.e., initial cutscore). The second pertains to the correspondence between final cutscores and the performance level descriptions, which judges used to base their judgments on, in the standard setting study.

5.9.1 Grammar Subtest

5.9.1.1 Distribution of Judges' Ratings of Essential Items

The judges' set of essential items for each criterion point are presented in Table 5.69. From the table it can be seen that the standard setting judges' selected a different number of items for each criterion point. Of the 22 judges, 8 (Judges 3, 4, 5, 6, 14, 17, 20, and 22) indicated that all the test items were essential for examinees to know in order to be exempted whereas 7 of the judges (Judge 2, 9, 10, 11, 16, 18, and 19) indicated that all items were essential in order to achieve even the third criterion point (in order to be placed in Level 4 of the English language support courses). Judges 9 and 16, on the other hand, selected quite a substantial number of items to represent the first criterion point (23 items). Judge 8 had five missing responses while Judges 15 and 21 had one missing response each.

On the whole, the largest number of items selected is for Criterion point 1; the second largest is for Criterion points 2 and 3. Criterion point 4 has the least number of selected items. This suggests that the judges' have relatively high expectations of the grammar elements that examinees are required to master. As far as individual judges are concerned, Judges 9 and 16 selected more items for Criterion point 1 (57.5%) as compared with the other judges. Judge 18 selected most of the items for Criterion point 2 (52.5%). Judge 6 selected the least number of items for Criterion point 1 (7.5%) and the most for Criterion point 3 (55%). Judge 21 selected the largest

number of items for Criterion point 4 (40%). Judges 1 and 3 have a somewhat fair distribution across the four criterion points. Note that only very few items were not selected. This suggests that judges on the whole expect examinees to get almost all items correct in order to be exempted from the English language support courses.

Table 5.69: Distribution of Grammar Items across Criterion Point by Individual Judges

Judge	Criterion Point					Total Items	Mode
	1	2	3	4	Not Selected		
J1	9 (22.5%)	13 (32.5%)	7 (17.5%)	10 (25.0%)	1 (2.5%)	40	2
J2	12 (30.0%)	18 (45.0%)	10 (25.0%)	-	-	40	2
J3	8 (20.0%)	9 (22.0%)	13 (32.5%)	10 (25.0%)	-	40	3
J4	14 (35.0%)	9 (22.5%)	12 (30.0%)	5 (12.5%)	-	40	1
J5	15 (37.5%)	5 (12.5%)	16 (40.0%)	4 (10.0%)	-	40	3
J6	3 (7.5%)	12 (30.0%)	22 (55.0%)	3 (7.5%)	-	40	3
J7	13 (32.5%)	15 (37.5%)	9 (22.5%)	2 (5.0%)	1 (2.5%)	40	2
J8	6 (15.0%)	9 (22.5%)	15 (37.5%)	5 (12.5%)	-	35	3
J9	23 (57.5%)	12 (30.0%)	5 (12.5%)	-	-	40	1
J10	17 (42.5%)	11 (27.5%)	12 (30.0%)	-	-	40	1
J11	19 (47.5%)	10 (25.0%)	11 (27.5%)	-	-	40	1
J12	17 (42.5%)	7 (17.5%)	8 (20.0%)	5 (7.5%)	3 (7.5%)	40	1
J13	13 (32.5%)	12 (30.0%)	7 (17.5%)	7 (17.5%)	1 (2.5%)	40	1
J14	5 (12.5%)	12 (30.0%)	14 (35.0%)	9 (22.5%)	-	40	3
J15	11 (27.5%)	15 (37.5%)	12 (30.0%)	1 (2.5%)	-	39	2
J16	23 (57.5%)	15 (37.5%)	2 (5.0%)	-	-	40	1
J17	18 (45.0%)	18 (45.0%)	3 (7.5%)	1 (2.5%)	-	40	1
J18	12 (30.0%)	21 (52.5%)	7 (17.5%)	-	-	40	2
J19	15 (37.5%)	16 (40.0%)	9 (22.5%)	-	-	40	2
J20	14 (35.0%)	13 (32.5%)	12 (30.0%)	1 (2.5%)	-	40	1
J21	8 (20.0%)	5 (12.5%)	10 (25.0%)	16 (40.0%)	-	39	4
J22	11 (27.5%)	8 (20.0%)	16 (40.0%)	5 (12.5%)	-	40	3

Table 5.70 gives the frequency of judge selection of individual items across the four criterion points. Some ‘easy’ items (as indicated by their empirical item calibrations) were found to elicit general agreement among judges as regard their placement. Items 2 (adjective – degree of comparison), 5 (Subject-verb agreement), 11 (preposition – time), 21 (Simple past tense), and 22 (preposition – location), indicate the least amount of variability as evidenced by their standard deviations.

On the other hand, item 7 (conjunction – ‘unless’) indicates the largest variability (*SD* – 1.71) followed by item 28 (modal – showing advice) with a standard deviation of 1.4. Other items that indicated standard deviations of above 1.0 are items 9, 18, 19, 29, and 31. Note that very few items were not selected by the standard setting judges as essential items.

Table 5.70: Frequency of Judges' Selection of Grammar Items by Criterion Point

Items	Criterion point					Mode	Standard Deviation
	1	2	3	4	Not Selected		
G1	6	12	4	0	-	2	0.68
G2	17	4	1	0	-	1	0.55
G3	3	11	8	0	-	2	0.69
G4	2	8	11	1	-	3	0.74
G5	18	2	2	0	-	1	0.63
G6	4	5	12	1	-	3	0.86
G7	4	6	6	5	1	2[a]	1.71
G8	2	11	8	1	-	2	0.73
G9	1	8	5	5	1	2	1.17
G10	1	5	8	6	1	3	0.97
G11	21	0	1	0	-	1	0.43
G12	7	8	6	1	-	2	0.90
G13	11	7	3	1	-	1	0.88
G14	11	8	2	0	-	1	0.68
G15	9	9	2	2	-	1[a]	0.94
G16	2	7	12	1	-	3	0.74
G17	13	6	3	0	-	1	0.74
G18	9	9	1	3	-	1[a]	1.02
G19	8	8	2	4	-	1[a]	1.11
G20	12	7	3	0	-	1	0.73
G21	21	1	0	0	-	1	0.21
G22	16	6	0	0	-	1	0.46
G23	8	6	7	1	-	1	0.95
G24	3	10	7	2	-	2	0.85
G25	19	1	1	1	-	1	0.77
G26	5	9	8	0	-	2	0.77
G27	3	9	9	1	-	2 [a]	0.79
G28	3	9	9	1	-	1[a]	1.4
G29	15	3	2	2	-	1	1.01
G30	1	9	7	5	-	2	0.88
G31	3	9	5	5	-	2	1.01
G32	3	11	8	0	-	3	0.69
G33	1	4	11	6	-	3	0.82
G34	1	6	8	6	1	3	0.98
G35	2	5	8	7	-	3	0.97
G36	11	7	3	0	-	1	0.74
G37	6	11	4	0	-	2	0.70
G38	1	7	11	2	-	3	0.73
G39	1	6	12	2	-	3	0.72
G40	0	4	13	4	-	3	0.63

[a] Multiple modes

5.9.1.2 Descriptive Statistics and Judge Variability

In examining judge variability several aspects are of interest. The first pertains to the relative position of the individual mean estimates established for each judge. These mean estimates are the means of the individual judges' respective distribution of essential items.

The second relates to the overall mean of the judges' distribution of mean estimates (i.e., the judges' individual mean estimates of essential items). The overall mean, which is the average of the judges' individual mean estimates, marks the criterion point that is used as the initial cutscore before it is adjusted for error of measurement to arrive at the final cutscore.

The third aspect involves the minimum and maximum mean estimates of judges' distributions for the four final criterion points. It is expected that the minimum and maximum mean estimates of judges' distributions advance in an increasing trend from the lowest criterion point (i.e., Criterion Point 1) to the highest criterion point (i.e., Criterion Point 4).

The fourth relates to the variability of judges' mean estimates. Variability of judges' mean estimates shows the extent to which judges' individual estimates are dispersed or spread out relative to the mean of its distribution. This is important as it demonstrates how much judges differ from one another. The statistic that is used to investigate variability is the standard deviation. A small standard deviation shows that judges generally do not differ much with each other while a large standard deviation shows substantial disparity in judgment.

5.9.1.2.1 Distribution of Judges' Mean Estimates of Essential Items

Figure 5.44 gives the distribution of judges' mean estimates of essential items for the four criterion points. From the figure, the location of each judge's mean estimate in relation to other judges can be seen. It is obvious that Judge 16

overestimated Criterion Point 3 whereas Judge 7 overestimated Criterion Point 4. Judges 13, 17 and 22 are found to underestimate Criterion Point 4.

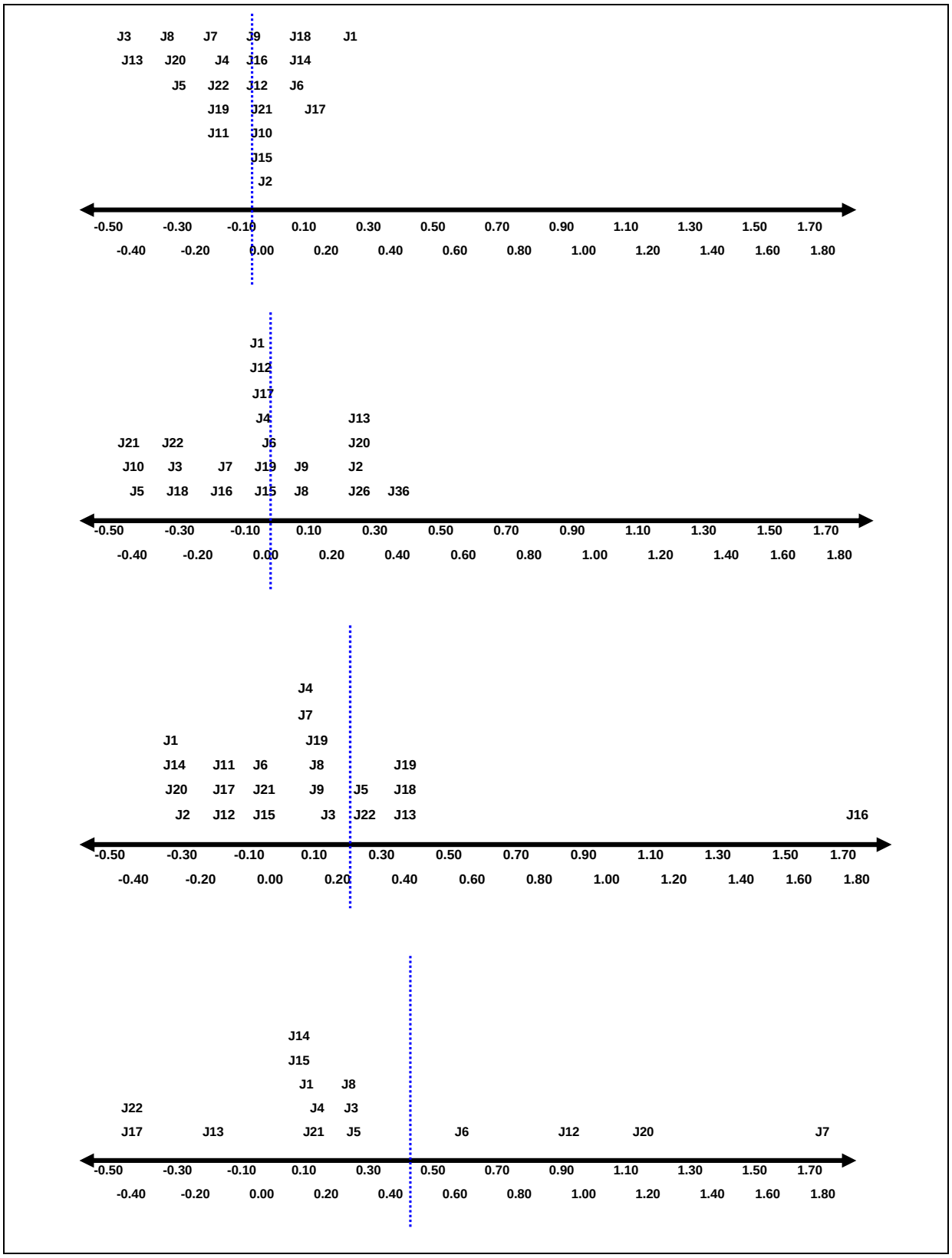

Figure 5.44: Distribution of Judges' Mean Estimates for the Four Criterion Points (Grammar Subtest)

Table 5.71 gives the descriptive statistics pertaining to the judgment of essential items for the four criterion points of the grammar subtest. The means of judges' distribution for the criterion points are not clearly distinguished as the difference between each mean is very small (between 0.05 to 0.24 logit). The minimum judge estimates of essential items for the four criterion points also indicate no real difference.

In addition, these estimates show a disordering. The minimum mean estimate of judges' distribution for Criterion Point 2 (-0.49 logit) is smaller than the minimum mean estimate for Criterion Point 1 (-0.47 logit). The minimum mean estimate for Criterion Points 3 and 4 show a similar disordering. However this trend is not seen with the maximum mean estimates.

Table 5.71: Descriptive Statistics of Judges' Mean Estimates of Essential Items (Grammar Subtest)

Criterion Point	Minimum	Maximum	Mean	Std Deviation
Criterion Point 4	-0.41	1.80	0.37	0.59
Criterion Point 3	-0.31	1.80	0.13	0.45
Criterion Point 2	-0.49	+0.36	-0.04	0.24
Criterion Point 1	-0.47	+0.25	-0.09	0.16

5.9.1.2.2 Judge Variability

As regards variability of judges' estimates, Criterion Point 1 shows the least amount of variation (*SD* = 0.16) followed by Criterion Point 2 (*SD* = 0.24) (Refer to Table 5.71). The largest variability is for Criterion Point 4 (*SD* = 0.59). These values are considered acceptable as Stone (1996) had reported that standard deviations of between 0.20 and 0.50 have been found to be uncommon in standard setting studies.

The boxplots in Figure 5.45 show the distribution of judges' mean estimates for the four criterion points. The boxplot for Criterion Point 2 indicates that there are no outliers or extreme cases. Criterion Point 1 shows 3 judges who are outliers; Criterion Point 3 has one extreme case; and Criterion Point 4 has one outlier and one extreme case (which could have contributed to the relatively large standard deviation for this criterion point).

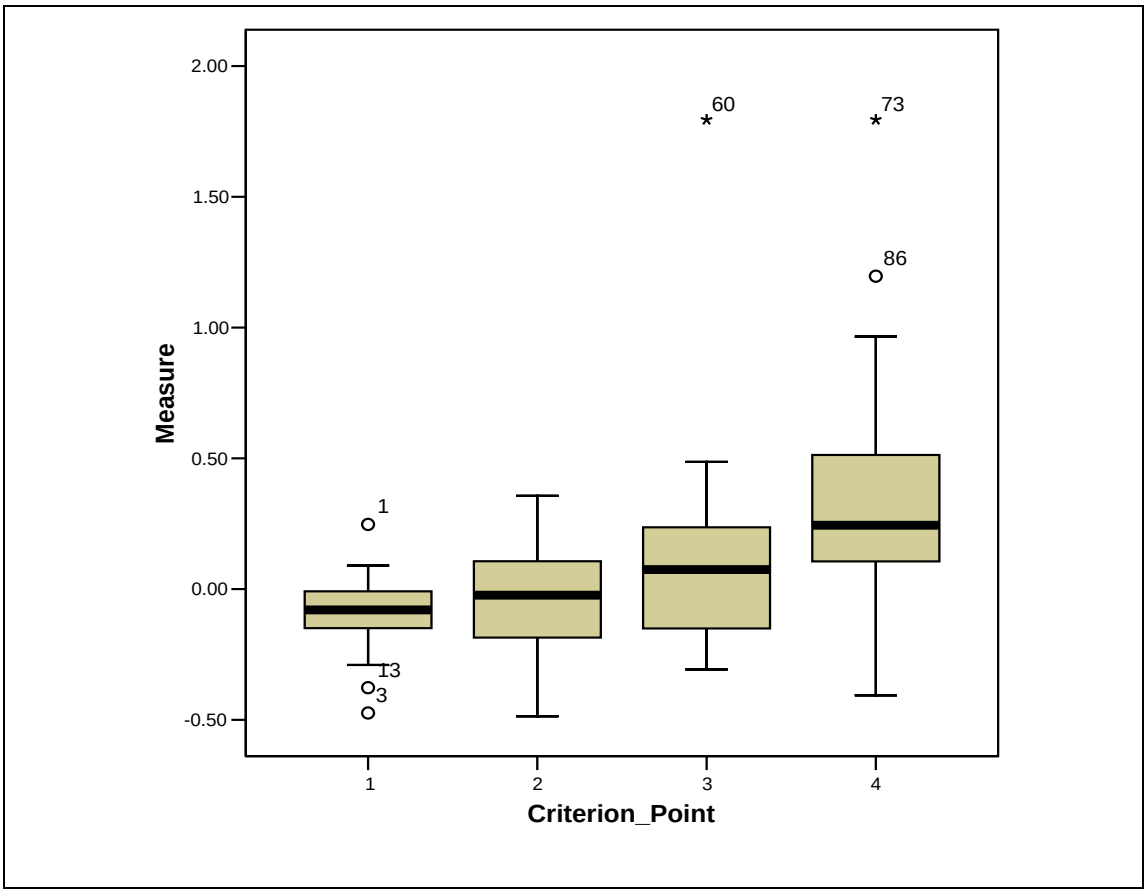

Figure 5.45: Boxplots of Judges' Mean Estimates for the Four Criterion Points (Grammar Subtest)

The following figure (Figure 5.46) demonstrates judges' estimation of the criterion points. The four lines indicate the four criterion points. The crossing of the lines gives a clear indication that judges' estimation of the four criterion points is rather haphazard.

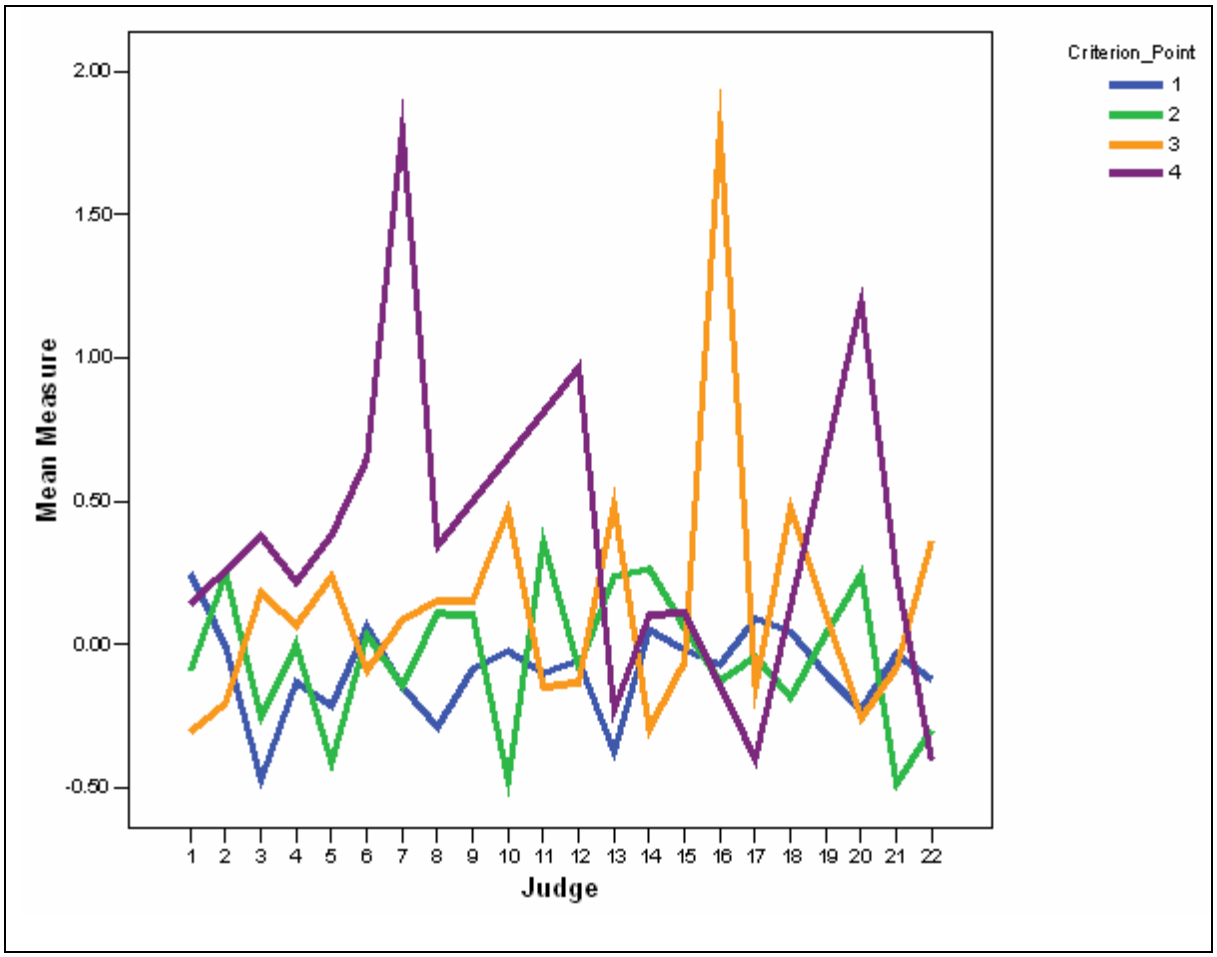

Figure 5.46: Judges' Estimation of the Four Criterion Points (Grammar Subtest)

In order to examine individual judges' estimation of the criterion points, line graphs for individual judges were plotted. The set of line graphs in Figure 5.47 show that only 6 judges (Judges 3, 4, 7, 8, 9, and 19) have made reasonable estimates. Judge 17 shows a reverse trend for all the four criterion points whereas Judges 5, 10, 16, 18, and 21 overestimated Criterion Point 1, underestimated Criterion Point 2 but made reasonable estimates for Criterion Points 3 and 4.

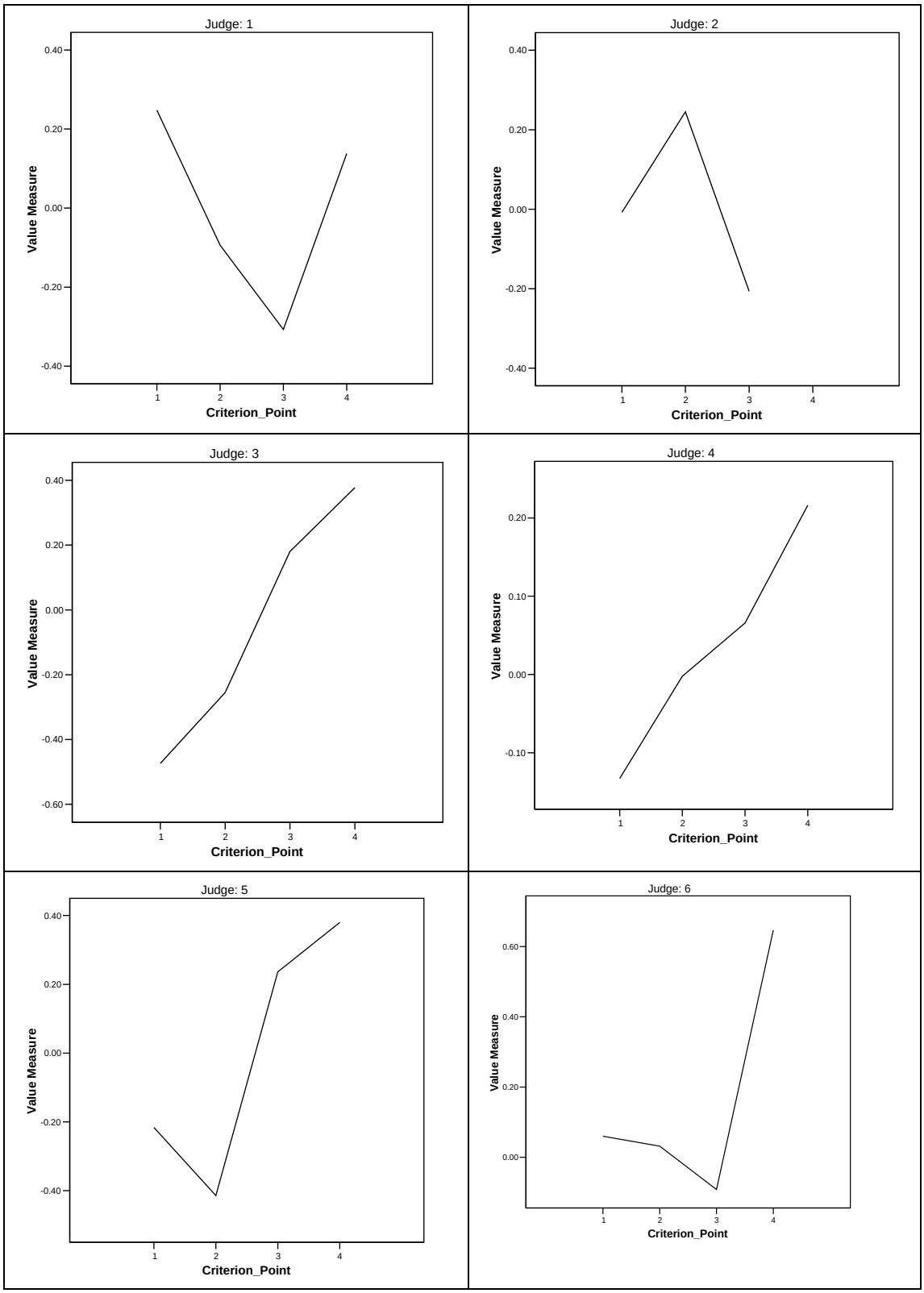

Figure 5.47: Individual Judges' Estimation of the Four Criterion Points (Grammar Subtest)

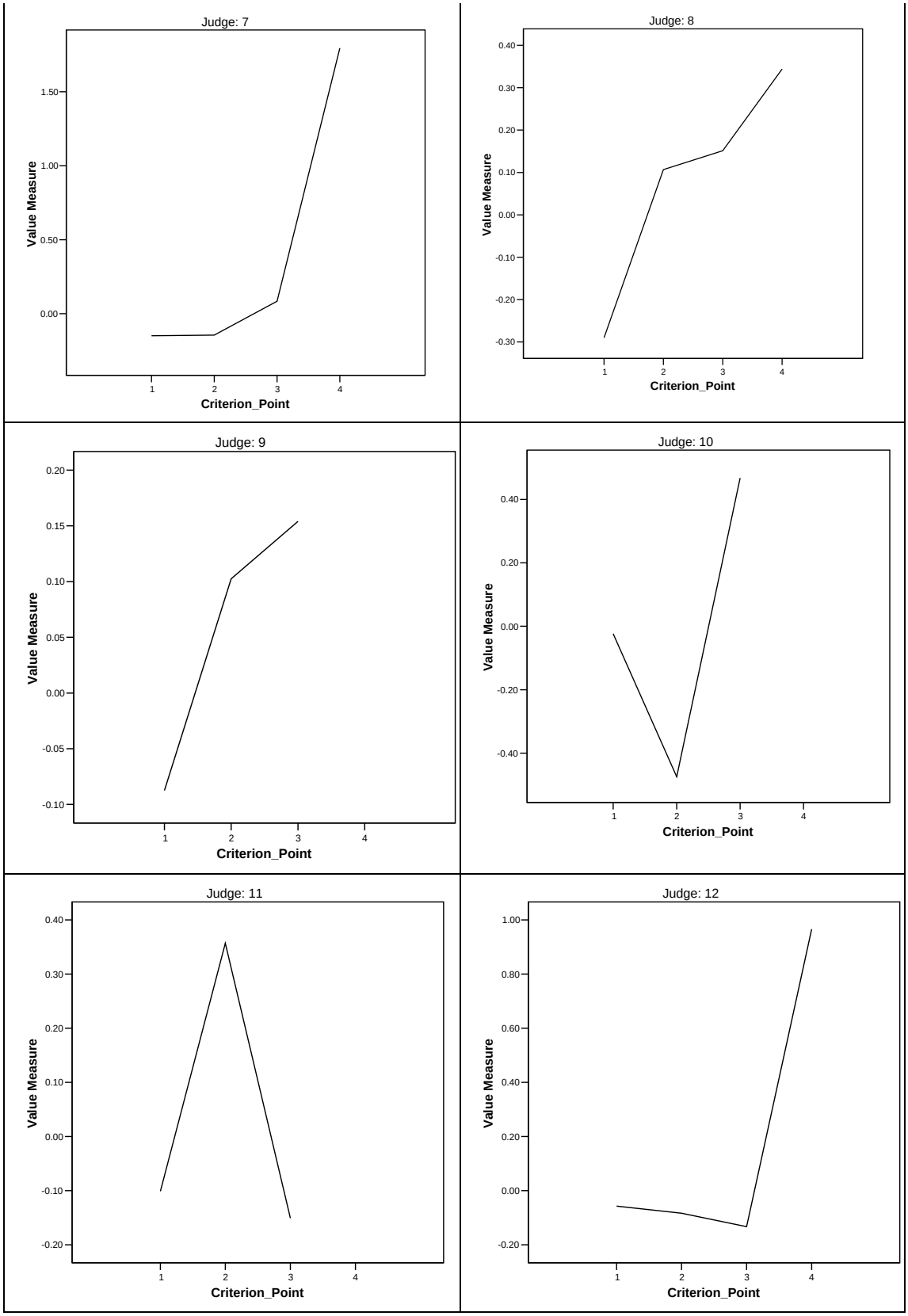

Figure 5.47.... continued.

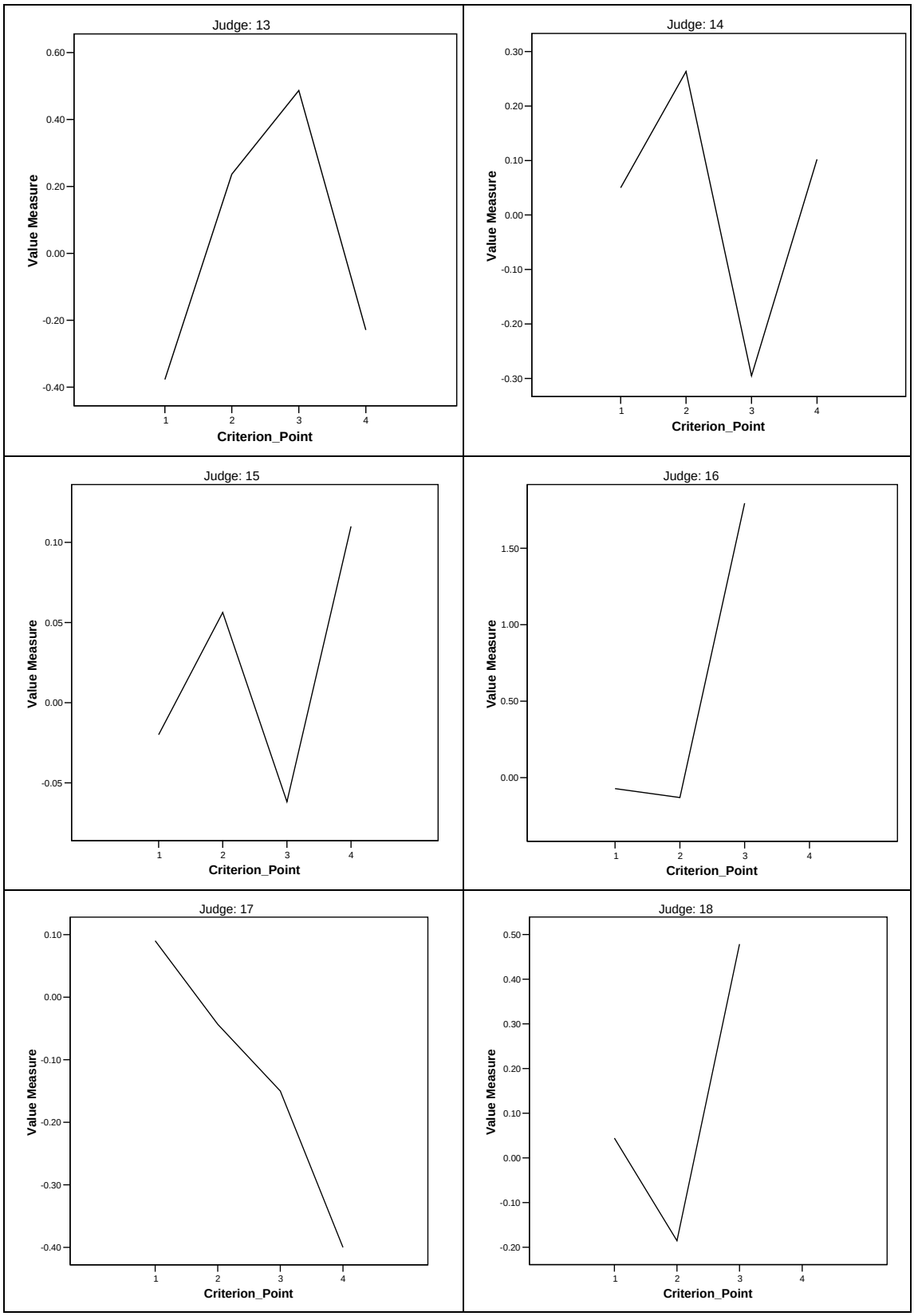

Figure 5.47.... continued.

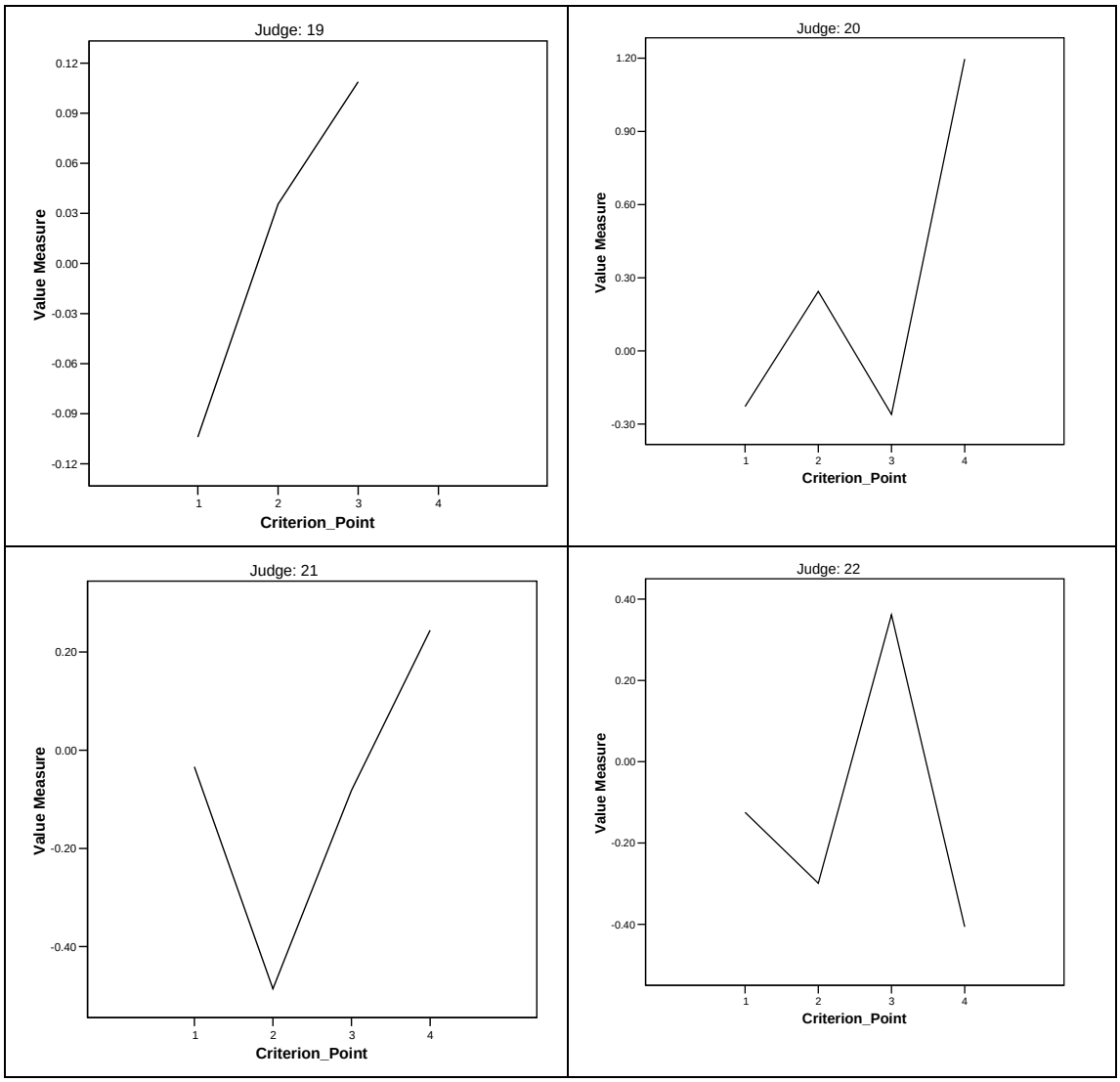

Figure 5.47.... continued.

5.9.1.3 Facets analysis

As mentioned in Chapter 4, Section 4.5.2.2, interjudge agreement was also investigated using a procedure recommended by Linacre (personal communication, June 23, 2005). Figure 5.48 presents the calibrations of judges, criterion points and items from this analysis. The first column on the left is the logit scale followed by the distribution of standard setting judges, criterion points and test items. From the figure, judges appear to differ from one another quite substantially as judge severity ranges about 3 logits (-1.5 logits to +1.8 logits). Item distribution is similar to the one in the test administration as item estimates in this analysis were anchored to the values derived in

the test administration analysis. As regards the criterion points, Criterion Points 3 and 4 are overestimated in relation to the item distribution.

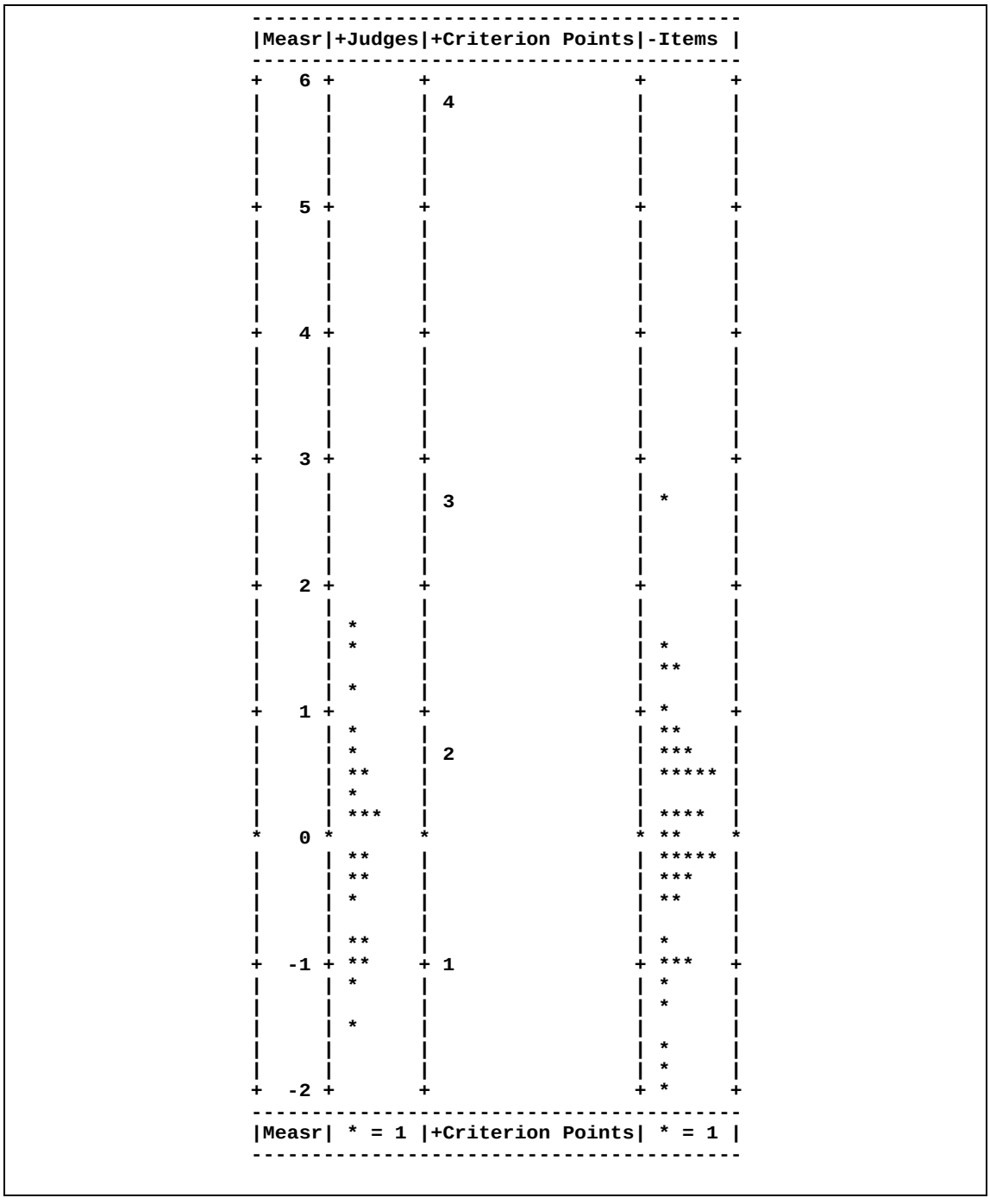

Figure 5.48: Calibrations of Judge Severity, Criterion Points and Test Items (Grammar Subtest)

5.9.1.3.1 Interrater Agreement

It is evident that there is variation in judges' perception of essential items. Judges' severity varies about 3 logits (-1.54 to +1.68 logits) (Table 5.73). Judges with high logit estimates are more severe in their estimation, in the sense that they have selected more items that are considered essential for low criterion points. For example, Judge 16 (severity measure: +1.68 logits) has actually selected 57.5 % of the items to represent Criterion Point 1 and 37.5% to represent Criterion Point 2 (Table 5.72). Likewise, Judge 9 (severity measure, +1.45 logits) displays a similar judging behaviour. The most lenient judge is Judge 21 (severity measure, -1.54 logits) followed by Judge 14 (severity measure, -1.08). The distribution of essential items for these judges indicates that a small number of items were selected to represent the lower criterion points. More items were selected by these judges to represent the higher criterion points.

Table 5.72: Distribution of Items across Criterion Point for Most Severe and Most Lenient Judges (Grammar Subtest)

Judge	Criterion Point				Total	Mode
	1	2	3	4	Items	
J9	23 (57.5%)	12 (30.0%)	5 (12.5%)	-	40	1
J16	23 (57.5%)	15 (37.5%)	2 (5.0%)	-	40	1
J14	5 (12.5%)	12 (30.0%)	14 (35.0%)	9 (22.5%)	40	3
J21	8 (20.0%)	5 (12.5%)	10 (25.0%)	16 (40.0%)	39	4

The statistical significance of judge variability is also examined using several indexes (Table 5.73). The judge separation index of 3.31 and the chi-square value of 245.8 with 21 df, significant at p <.01 indicate that judges consistently differ from one another in overall severity (Table 5.73). The observed number of exact agreement for all judges is relatively high: 37893 (83.3%) out of a total of 45465 exact agreement opportunities.